AISB 2011

Machine Consciousness

Editors:
Dimitar Kazakov &
George Tsoulas

THE UNIVERSITY of York

Foreword from the Convention Chairs

The AISB'11 call for symposium proposals particularly encouraged events drawing more strongly on the cognitive science aspect of the AISB remit. The result is a coherent programme with a very strong interdisciplinary character, which is also matched in the choice of plenary speakers. The three symposia looking at the interaction between Computing and Philosophy, the prospect of machine consciousness and the quest for a new, comprehensive intelligence test, form a coherent unit where the eternal questions of who we are and what makes us so are asked from a dual Human-Machine perspective. The Symposia on Active Vision, Computational Models of Cognitive Development and Human Memory for Artificial Agents demonstrate how better understanding of the nature and basis of cognitive processes can advance work on Artificial Intelligence and, inversely, how computational models of these processes can help better to understand them. The prominent multi-agent design and modelling paradigm links the Symposium on Social Networks and Multi-agent Systems with the one on AI and Games. Finally, the Symposium on Learning Language Models from Multilingual Corpora, which brings together some of the first attempts in this area, can also be seen through the prism of such a general notion in Philosophy and Linguistics as semiosis, and the dual role of sign and interpretant that text plays in translations.

We are delighted that after another ten successful years in its long history, the AISB convention is returning to the University of York. The 2011 convention takes place on the brand-new Heslington East campus, the result of a multi-million pound expansion that is now the new home of the Department of Computer Science, and hosts the Excellence Hub for Yorkshire and Humber, a new incubator for interdisciplinary research and interaction between academia and industry. The last few years have seen a strong involvement of the Computer Science Department in such interdisciplinary collaboration through the York Centre for Complex Systems Analysis (YCCSA), and we hope that this convention will provide a boost for more synergy between York departments, with other institutions conducting AI-related research in the region, and beyond. As the programme shows, we have also made an effort to promote cooperation with industry and use the convention to support school outreach. The convention format makes it perfect for establishing dialogue and collaboration in new areas of research, as well as across disciplines, and we hope that this year, it will play again this role to the full. We want to thank everyone who has contributed to it or otherwise made this event possible and wish all participants a fruitful and enjoyable time in York.

Dimitar Kazakov and George Tsoulas

Machine Consciousness 2011: Self, Integration and Explanation

Proceedings of a symposium at the AISB'11 Convention

4-7 April 2011, York, United Kingdom

Edited by Ron Chrisley, Rob Clowes and Steve Torrance

Published by the UK Society for the Study of
Artificial Intelligence and Simulation of Behaviour

Introduction

Rob Clowes* **, Steve Torrance* and Ron Chrisley*

*Centre for Research in Cognitive Science (COGS), University of Sussex, Falmer, Sussex BN1 9QH UK
**Institute for the Philosophy of Language, New University of Lisbon (Universidade Nova de Lisboa)
Email: {robertc, stevet, ronc} <at> sussex.ac.uk

Machine Consciousness as a field: Its purpose, recent history and present

The field of research called Machine Consciousness (MC) – or if you prefer artificial or synthetic consciousness - might date its existence from almost exactly 10 years before this workshop – actually May 2001 – when the Swartz Foundation organised a workshop called 'Can a machine be conscious?' The detailed early history as a field is detailed in Holland´s 2003 editorial introduction [1] to the first journal special issue – for the *Journal of Consciousness Studies* – on Machine Consciousness. It is instructive to look back at this special issue, and Holland´s introduction, to see both some of the continuities of the field and some of what has changed.

An ongoing problematic is that there is still no firm agreement on what the aim of the field is. For many researchers, certainly from the first wave of papers, the idea was to build actually conscious machines. Cotterill, in the original JCS special issue, articulated this well when he wrote that "a step toward realizing the long-cherished dream of creating *Homo silicens*: consciousness in a computer" [2].

Others have different ideas about the purpose and utility of the field. It might be that its main purpose is to help us understand the nature of consciousness by modelling naturally occurring consciousness? [3] Or perhaps to help us specify the content of consciousness? [4]. In part – but only in part – these divisions can be reduced to whether we are speaking of "weak" or "strong" MC research [3, 5].

It is a point of historical debate as to whether the field of MC should be considered as broadly continuous with its parent field of artificial intelligence or whether it should be considered as a bold departure. Perhaps while Artificial Intelligence was concerned merely with intelligence, which turned out to be multiform, MC has focused in on a much more specific cluster of questions about how one might build, or use simulations or robotic models to understand, minds that there is something it is to be like. This point turns on whether we suppose that consciousness is co-extensive with intelligence of a certain order, or whether one could build a machine of high – perhaps human level – intelligence which nevertheless should not be considered conscious. Questions on this point are far from resolved, and are taken up at our conference by Torrance in his paper: *Would a super-intelligent AI necessarily be (super-) conscious?* (this volume). Torrance poses many tough questions to MC researchers and maps out a series of positions one might take on the matter. Recalling Cotterill´s notion of the "long-cherished dream" it is an interesting to consider whether machine consciousness has made progress over the last ten years either toward building actually conscious machines – and here we must note that some would not consider this progress, e.g. [6] – or indeed whether we have made progress toward understanding (natural) consciousness.

The present workshop on MC follows on from two others organised as part of AISB annual Conventions, at Hertfordshire [7] and Bristol [8]. A special issue of the *Journal of Consciousness Studies* followed, with papers based largely on contributions to those workshops [9] . Being asked once again – five years after the last one – to organise a workshop on MC for an AISB Annual Convention gave us the opportunity to consider how the field has changed in the last five years and more importantly, perhaps, where it might be going. We should note that in the *JCS* special issue we identified two core themes that many of the papers addressed: Imagination and Embodiment. It remains to be seen how dominant these themes are this time round. The Imagination theme was certainly seen previously by many researchers to be central to understanding consciousness and to its implementation in machines [10],[11],[12],[13].

The other focus, embodiment, centred on the proposition that, for a machine to be actually conscious it was not enough for that machine to be implemented *in silico* but that it also had to be embodied, in some sense, and dynamically interacting in the world. Yet – as several papers pointed out at the time, especially [14] and [5], quite what the right sort of embodiment should be remained a difficult question. These themes are by no means *passé* as we shall see below. They may well also be touched on by our invited speaker Murray Shanahan, whose new book *Embodiment and the inner life: Cognition and Consciousness in the Space of Possible Minds* [15] considers these themes.

Ten years down the line from the Swartz foundation workshop, and five years after the previous AISB conference, the field appears to have made great strides forward, with the creation of a new journal – the *International Journal of Machine Consciousness* – and several rather advanced research programmes: a number of these are represented in papers in the present Proceedings. MC has even taken a central role in one of the prominent introductions to the field [16]. Yet it seems worthwhile to try to ask: what progress has MC actually

made over this period? Was its principal advance to further elaborate existing theories? Or was it seeking to pioneer new ground empirically and/or theoretically? It also seemed worth discussing how MC fitted into a wider discussion of the place of consciousness in nature, and how machine consciousness research relates to the programmes of cognitive psychology and neuroscience, as well as to "mainstream" AI. It was largely because it gave us an opportunity to put these questions on the agenda that we decided it was right to organize another AISB workshop.

Here are some of the themes that seem to be emerging for discussion in this workshop, on the basis of contributions received.

Global Workspace as a Model for MC

Perhaps one of the most advanced frameworks so far articulated is Franklin et al´s LIDA. The LIDA model is explicitly grounded in one of the most widely recognised theories of consciousness, namely Baars' Global Workspace Theory [17, 18] [see also 15]. Ramamurthy and Franklin's contribution for this volume appears to cohere well with that suggested for *weak* machine consciousness approach. Rather than making strong claims that IDA is actually phenomenally conscious, it helps elaborate a more detailed and articulated version of the Global Workspace Theory – and might also help to show the theory's relation with other theoretical approaches. The authors note, for instance, that the LIDA model implements aspects of "situated cognition, working memory, memory by affordances, long-term working memory, Sloman´s H-CogAff and transient episodic memory." Demonstrating that aspects of such theories can be implemented within a single cognitive model is unquestionably an impressive feat. It is less clear what the explanatory status of such models is and indeed the question of what explanatory role of MC research more generally is a question we hope to probe further at the conference.

Raizer, Paraense and Gudwin et al (this volume) also provide a contribution that is inspired by the Baars/Franklin LIDA model. Their research focuses on building a variety of artificial creatures with varying cognitive architectures, each of which exemplifies a certain level or degree of consciousness (as defined in "consciousness scales" such as the one developed by Arrabales *et al.* in their ConsScale [19]. Raizer and colleagues base their models on the Baars-Franklin functional framework. They are proposing to produce a series of progressively more complex models of conscious agency, utilizing the conception of the different evolutionary stages represented in the human brain (reptilian, paleomammalian, neomammalian) developed by McLean [20].

Information Integration

Another major theme of papers at the conference is information integration. Information integration [21] has been one of the strongest candidates for an empirically testable notion of consciousness – and arguably one that synthetic approaches might directly address. It sets out to capture the intuition that the fine-grained nature of phenomenal consciousness can be regarded as both informative and explanatory. Unlike almost all other theories it makes strong predictions about which systems are conscious and in principle makes possible the development of straightforward measures of consciousness.

The paper from Aleksander and Beaton (this volume) looks more strictly at what the predictions of the existing information integration approach might make. Aleksander and Beaton argue that neither version of Tononi´s information integration theory ([22](Tononi 2008) provides an adequate account of how information is both integrated and differentiated within consciousness. They make concrete proposals to address this problem.

Gamez (this volume) looks at how a more developed version of the information theory might be tested, namely:

1. Select a system that is known or commonly agreed to be conscious.
2. Measure the information integration of the system.
3. Measure the consciousness of the system.
4. Identify correlations between information integration and consciousness.
5. Test predictions made by information integration about the consciousness of the system.

Of course this appears to presuppose an independent measure of consciousness, which is arguably what we are looking for information integration to do in the first place. Several of these do exist (see Arrabeles' and Gamez´own previous work). It should also be noted that it is not clear to what extent or how information integration approaches can help us determine the content of a particular experience (see the discussion of synthetic phenomenology, below). And even if information integration could be established to be a sufficient condition for consciousness, rather than only being a correlate in the presence of some further, enabling condition (such as embodiment, or situatedness – see the final section), there would still be room to question it as a necessary condition. That is, it appears that one could not infer, without further support, the absence of consciousness based on a low degree of information integration. Finally, there is reason to doubt some of the applications that Gamez envisages. For example, many deny that there is a distinct phenomenology for intentional states such as intentions. If they are right, then information integration-based mind reading would be of little use to the courts, at least along the lines Gamez proposes.

Despite the plausibility of the idea that information integration may be central to constituting some basic form of consciousness, many researchers seem convinced that human (and perhaps much animal) consciousness is structured by a fundamental division in representational space between self and non-self. Such a division might

play a role in constituting any sort of consciousness, or alternatively conferring a more elevated form of consciousness that we might call subjectivity. This conceptual division, as seen in theoretical models proposed by Damasio and Metzinger, has inspired several machine consciousness researchers to investigate self-models.

Instantiating Self Models

The idea that self-modelling systems might be central to at least certain orders of consciousness is one which has been central since the first volume on machine consciousness [23]. Holland´s work pioneered the self-model approach to machine consciousness but linked self-models directly to the problem of controlling and maintaining a motile body (an approach pioneered theoretically by Damasio and Metzinger). Since Holland´s original work there has been a plurality of suggestions both on how self-modelling might feature in consciousness research and what sorts of models might be most useful. The idea also figures in a number of contributions to this workshop. Kinouchi *et al,* Ramamurthy and Franklin, Sanz *et al*, and Taylor all use the idea of self-modelling in different ways. It should be noted however that in Metzinger´s work – which was one source of inspiration for Holland´s models – having a self was not even a necessary component of being conscious, rather it figured in a particular advanced configuration of consciousness: subjectivity.

According to Damasio´s approach to self, the core stability necessary for an ongoing self is based in the body's need to maintain life within the rather narrow confines of viability [24]. His basic typology – core self, minimal self and extended self – has been taken as an eminent target for modelling, although it is difficult to say how much can be learnt from these models in the absence of a body. (Of course this once again raises the question of what sort of body is needed? The problem of embodiment is never far away.).

There are several attempts to develop versions of this approach in evidence for the 2011 conference. The first of these, by Kinouchi *et al*. (this volume), offers a model of 'primitive consciousness' from which self emerges in a "logical space". The design is based on an autonomous system of six interacting modules: a perception module, an integration module, a motor control module, an episodic memory module and a working memory module. Because at this stage the authors have designed only an abstract model yet to be implemented, it is difficult to say whether the abstract informational structure of the "self-model" will be tied to any robotic instantiation or embodiment. Certainly some other researchers appear not to think this is necessary to demonstrate useful results.

One of the more interesting attempts to use self-modelling approach inspired by Damasio can be found in the paper by Ramamurthy and Franklin. This work extends the much discussed LIDA system which - as we have seen - is articulated within a Baars GWT framework. The paper presented here seeks to marry this approach to the ideas of self-models as understood by Damasio. One perhaps surprising aspect of this, given that Damasio´s own approach is closely tied to *biological* systems, is that LIDA has a core self which is entirely modelled *in silico;* and, in contrast to the robotic work we have just discussed, the "proto-self" appears to be based on no robotic embodiment. How compatible this approach really is with Damasio´s model is therefore a matter for debate. For Damasio the proto-self was based firmly on the processes in a biological body. Although an *in silico* model can in some sense clearly be produced, it remains unclear what the explanatory significance of such models are. Are they supposed to act as tests of theory? Or as ways of elaborating the theory in more detail? When is a self-model the right sort of self-model? Answering the latter question seems some way off although perhaps the models discussed at this workshop will help clarify this issue.

Phenomenology and functionality

Closely related to the strong/weak MC issue mentioned earlier on is the relation between functional and phenomenal consciousness. Since Block introduced this, or a closely-related, distinction [25] there has been a lot of discussion of how far MC research can accommodate such a distinction, and to what extent MC research is able to do justice to our full phenomenology, as opposed to certain limited functional aspects of our consciousness [26]. There are a number of positions that can be taken here, of which the main ones are: (1) that the functional/phenomenal consciousness distinction is not a valid one; (2) that the distinction is valid, but that MC systems could be produced with both kinds of consciousness; and (3) that MC can only make progress on modelling or instantiating functional consciousness. Many MC researchers defend position (1), but some defend (2), and some will even agree that (3) may be the case.

Defending position (3) is one way of articulating the view that strong MC is not achievable. In his contribution, Torrance (this volume) makes a number of points in support of (3), in particular arguing that ascribing phenomenal, as opposed to functional, consciousness to a being (natural or artificial), has certain ethical ramifications: we would expect that conscious artificial agents should be accorded certain moral rights (and even responsibilities) which we would not be so ready to accord to beings that 'only' had functional consciousness. However he agrees that this is still very much an open issue, and he points out that this puts some responsibilities in turn upon MC researchers, especially in relation to possible 'superintelligent' beings, as predicted by singularity theorists, as their super-human cognitive abilities may imply that they also have super-human kinds of consciousness, and hence, perhaps, super-human ethical claims.

Taylor's paper (this volume) also argues for a clear distinction between functional and phenomenal consciousness. He discusses his CODAM system (Corollary Discharge of Attention Model – see [27]) in relation to different levels in which attention may be manifested in the control systems of kinds of agents. As

with Raizer and colleagues, he takes an evolutionary perspective, and argues in some detail, that phenomenal consciousness, or "phenomenology", is dependent upon their being a "corollary discharge attention control signal", which ensures that there is, in effect, an "Inner self" which is the "owner" of the experience. It is this complex attentional control that distinguishes us from "zombies". As his system is one which is able to implemented as an MC system he seems, in effect, to be arguing in favour of proposition (2) above.

Sanz, Hernández and Sánchez (this volume) also consider the phenomenology of a system or agent, but from the somewhat different perspective of control system engineering. They see phenomenology as closely associated with the possession of intrinsic goals or teleology: a control system (such as one that runs an industrial plant or an aeroplane) may be considered as having its own phenomenology to the extent to which it takes the goals built-into its requirements-specification as its own goals (i.e. to the extent that the system understands what its users wish it to do, and integrates that understanding in a self-model that continuously updates itself in relation to changing conditions of perception and action). Sanz and colleagues support the adoption of a Dennettian heterophenomenological approach to engineering such self-awareness and intrinsic goal-ownership into systems. So it looks as though their position may be closer to (1) above, implicitly suggesting a continuity between functionality and phenomenology.

Synthetic phenomenology is the use of artefacts (such as computers and robots) to assist in the specification of experiential states. Chella and Gaglio (this volume) argue that previous approaches to synthetic phenomenology have been hindered by an emphasis on dimensionality reduction, and that if anything, such compression may have the undesirable result of eliminating phenomenology altogether. They propose an architecture for synthetic phenomenology, and detail an application of this architecture in a robot vision system, that uses a high-dimensional buffer to retain the richness of visual experience. The most explicit connection to phenomenology comes at the end of section 4, where they assert that the high-dimensional buffer will allow the robot to give appropriate answers to phenomenological questions (itself a kind of dimensionality reduction). What remains for future work is the development of some evaluative framework to compare the relative merits of low vs. high dimensional architectures in this area. A factor that will have to be taken into account in such a framework is the intelligibility of the phenomenological specification for the theorists using it, a factor which may require dimensionality-reduction and related data visualization techniques after all.

Chella and Gaglio note that their approach is influenced by Dennett's "multiple drafts" model of consciousness. Bryson (this volume), too, takes a broadly Dennettian approach in giving an account of the function of consciousness. She appeals to a range of studies of human and animal learning to argue that the function of consciousness is to limit the combinatorial complexity of certain learning situations. This enables her to characterize rather precisely the conditions under which a machine could be said to be conscious, and, perhaps more importantly for those interested in the technological relevance of MC research, when "adding conscious" would confer practical engineering and computational benefit.

Situatedness

Taking a philosophical approach that is quite distinct from the others in this volume, despite sharing with many of them an emphasis on mechanism and working implementation, Manzotti (this volume) places centre-stage the situatedness and time-dependence of consciousness. His starting point, based on the results of his previous research, is the conjecture that "the phenomenal experience of X might be nothing but the fact that X plays the twofold role of the cause of development and a current cause of behaviour." He offers a detailed cognitive architecture that permits him to give a clear interpretation of this idea. The key architectural notion is that of the states of a system becoming selectively sensitive to certain aspects of the environment in a history-dependent way, yielding the desired form of situatedness. Some philosophers will no doubt question whether such a simple causal account is consistent with the intentionality of at least some pehenomenal states, given the difficulties that causal approaches have had in accounting for, e.g., the possibility of error. Others may question his inference that causal connection implies constitution. And like other history-dependent accounts of mind, Manzotti's account seems to imply the possibility of the Swamp Man zombie: a creature that is (occurently) physically identical to you, but, since it does not have the right history, does not have any phenomenal consciousness at all. Those who wish to pursue these problems can turn to Manzotti's previous work for answers.

References

1. Holland, O., Editorial Introduction. *Journal of Consciousness Studies*, 2003. **10**(4): p. 1-6.
2. Cotterill, R., CyberChild A Simulation Test-Bed for Consciousness Studies. *Journal of Consciousness Studies*, 2003. **10**(4): p. 31-45.
3. Clowes, R.W. and A. Seth, Axioms, properties and criteria: Roles for synthesis in the science of consciousness. *Artificial Intelligence in Medicine*, 2008. **44**(2): p. 91-104.
4. Chrisley, R. and J. Parthemore, Synthetic phenomenology: Exploiting embodiment to specify the non-conceptual content of visual experience. *Journal of Consciousness Studies*, 2007. **14**(7): p. 44-58.
5. Torrance, S., Two Conceptions of Machine Phenomenality. *Journal of Consciousness Studies*, 2007. **14**(7): p. 154-166.
6. Metzinger, T., *Being No One: The Self-Model Theory of Subjectivity*. 2004: Bradford Books.
7. Chrisley, R., R.W. Clowes, and S. Torrance, Next-generation approaches to machine

consciousness, in *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*, R. Chrisley, R.W. Clowes, and S. Torrance, Editors. 2005.
8. Clowes, R.W., R. Chrisley, and S. Torrance, *Proceedings of the Symposium on Integrative Approaches to Machine Consciousness, AISB-06*. 2006, Bristol: University of Bristol.
9. Torrance, S., R. Clowes, and R. Chrisley, eds. *Machine Consciousness: Embodiment and Imagination,* Special Issue of *Journal of Consciousness Studies*. Vol. 14, Number 7. 2007.
10. Hesslow, G. and D.-A. Jirenhed, The inner world of a simple robot. *Journal of Consciousness Studies*, 2007.
11. Aleksander, I. and H. Morton, Why axiomatic models of being conscious? *Journal of Consciousness Studies,* 2007.
12. Haikonen, P.O.A., Essential issues of conscious machines. *Journal of Consciousness Studies*, 2007. **14**(7): p. 72-84.
13. Clowes, R.W., A Self-Regulation Model of Inner Speech and its Role in the Organisation of Human Conscious Experience. *Journal of Consciousness Studies*, 2007. **14**(7): p. 59-71.
14. Ziemke, T., The embodied self: Theories, hunches and robot models. *Journal of Consciousness Studies*, 2007. **14**(7): p. 167-179.
15. Shanahan, M., *Embodiment and the inner life: Cognition and Consciousness in the Space of Possible Minds*. 2010, Oxford: OUP.
16. Blackmore, S., *Consciousness: An Introduction. Second Edition*. 2010, London: Hodder Education.
17. Baars, B., *A cognitive theory of consciousness*. 1988, Cambridge: Cambridge University Press, Cambridge.
18. Baars, B., *In the Theater of Consciousness: The Workspace of the Mind*. 1996, New York: Oxford Univ Press.
19. Arrabales, R., A. Ledezma, and A. ConsScale, A Pragmatic Scale for Measuring the Level of Consciousness in Artificial Agents. *Journal of Consciousness Studies,* 17, 2010. **3**(4): p. 131-164.
20. MacLean, P.D., *The triune brain in evolution: Role in paleocerebral functions*. 1990: NY: Springer.
21. Tononi, G., Consciousness as Integrated Information: a Provisional Manifesto. *Biol. Bull*., 2008. **215**: p. 216-242.
22. Tononi, G. and O. Sporns, Measuring Integrated Information. *BMC Neuroscience*, 2003. **4**(31).
23. Holland, O. and R. Goodman, Robots With Internal Models A Route to Machine Consciousness? *Journal of Consciousness Studies*, 2003. **10**(4): p. 77-109.
24. Damasio, A.R., *The Feeling of What Happens: body, emotion and the making of consciousness*. 2000: Vintage.
25. Block, N., On a confusion about the function of consciousness. *Behavioral and Brain Sciences*, 1995. **18**(2): p. 227-247.
26. Franklin, S., IDA: A Conscious Artifact? *Journal of Consciousness Studies*, 2003. **10**(4): p. 47-66.
27. Taylor, J.G. The CODAM model of Attention and Consciousness. in *Neural Networks, 2003. Proceedings of the International Joint Conference on*. 2003: IEEE.

World-Related Integrated Information: Enactivist and Phenomenal Perspectives

Igor Aleksander[1] and Mike Beaton[2]

Abstract. *Information integration* is a measure, due to Tononi and co-researchers, of the capacity for dynamic neural networks to be in informational states which are unique and indivisible [1]. This is supposed to correspond to the intuitive 'feel' of a mental state: highly discriminative and yet fundamentally integrated. Recent versions of the theory include a definition of qualia which measures the geometric contribution of individual neural structures to the overall measure [2]. In this paper we examine these approaches from two philosophical perspectives, enactivism (externalism) and phenomenal states (internalism). We suggest that a promising enactivist response is to agree with Tononi that consciousness consists in integrated information, but to argue for a radical rethink about the nature of information itself. Using Cox's theorem, we argue that information is most naturally viewed as a three-place relation, involving a Bayesian-rational subject, the subject's evidence, and the world (as brought under the subject's evolving understanding). Therefore, to have (or gain) information is to behave in a certain (Bayesian-rational) way in response to evidence. As such, information only ever belongs to whole subjects (rationally behaving agents); and information is only 'in the brain' from the point of view of a theorist seeking to explain such behaviour. Moreover, rational behaviour (and hence information) will depend on brain, body and world – embodiment matters. From a phenomenal states perspective, we examine the way that internal states of a network can be not only unique and indivisible, but also reflect this coherence as it might exist in an external world. Extending previously published material [3], we propose that two systems could both score well on traditional integration measures where one had meaningful world representing states and the other did not. A model which involves iconic learning and depiction is discussed and tested in order to show how internal states can be about the world and how measures of integration influence this process. This retains some of the structure of Tononi's integration measurements but operates within sets of states of the world as filtered by receptors and repertoires of internal states achieved by depiction. This also suggests a formalisation of qualia which does not ignore world reflecting content and relates to internal states that aid the conscious organism's ability to act appropriately in the world of which it is conscious. Thus, a common theme emerges: Tononi has good intuitions about the necessary nature of consciousness, but his is not the only theory of experience able to do justice to these key intuitions. Furthermore, Tononi's theory has an apparent weakness, in that it treats conscious 'information' as something intrinsically meaningless (i.e. without any necessary connection to the world) whereas both the approaches canvassed here naturally relate experienced information to the world.

1 INTRODUCTION

Tononi's theory of integrated information starts from two unargued phenomenological intuitions. The first of these is that any given conscious experience is highly informative: that in having the experience, things seem one particular way rather than any one of an extremely large number of other ways that things might have appeared. This intuition seems correct: as Tononi points out, even an apparently simple experience, as of a *ganzfeld* of a pure colour, implicitly rules out many other experiences which I could have been having (such as being at the zoo, in the science museum, at my desk, etc. almost *ad infinitum*). Tononi's second intuition is that the information in experience is integrated – that all the separate distinctions made within one single experience are somehow unified. Again, this intuition seems sound. When I experience a blue chair, I am not somehow separately aware of the blueness and of the chair, but am aware of the combined whole. Indeed, when viewing an entire visual scene, my experience specifies a large number of different properties, at different locations, and these various distinctions are, once again, in some sense all integrated: all available to a single subject.

It is true that both these intuitions can be questioned. For instance some argue that a mode of pure experience exists, in which things do not seem to be any specific way at all [4]. This is incompatible with Tononi's claim that every experience at least implicitly contrasts itself with all other possible experiences. As for the second intuition, Metzinger, for one, has suggested that the unity of experience consists in the existence of a mental model as of unity, rather than in the existence of anything more fundamental which actually is unified [5].

However, our aim here is not to question Tononi's intuitions. In fact, we agree with them, but want to show that there are important aspects to consciousness which are overlooked in Tononi's further, formal development of Φ. We will argue that consciousness needs to be about the world, and that it needs to involve interaction with the world. We will show that it is possible to respect Tononi's two fundamental intuitions, *and* to capture these additional aspects, in at least two different theories of consciousness (one relatively externalist, and one more internalist). This weakens Tononi's strong claim that Φ corresponds directly to consciousness, and suggests that essential aspects of consciousness may be overlooked by Tononi's approach.

[1] Dept. of Electrical and Electronic Engineering, Imperial College, London SW7 3BT. Email: **i.aleksander@imperial.ac.uk**
[2] Dept. of Electrical and Electronic Engineering, Imperial College, London SW7 3BT. Email: **mjsbeaton@gmail.com**

2 THE PHOTODIODE AND THE CAMERA

To flesh out his two intuitions, Tononi contrasts the case of a conscious human being with the case of a photodiode, and with the case of a digital camera. Tononi points out that a photodiode can simply detect light above a certain intensity as "on", and below that intensity as "off". This is contrasted with the case of a human. When seeing a blank screen as either light or dark, the human is not just making the light/dark contrast which the photodiode can make, but is also seeing the screen as not being all the many, many other ways it could have been. Put another way, the human is not having all the many, many other distinct experiences which they could have had. Tononi suggests that this highlights a key difference between the photodiode and the human: the number of different states which can be distinguished. This is what he means by saying that consciousness is 'highly informative'.

Tononi then moves on to the case of a digital camera, in order to argue that merely 'generating a large amount of information' is not sufficient for consciousness. He considers a camera made from a million photodiodes. Such a camera can distinguish any one from among $2^{1,000,000}$ states, and there is no reason in principle not to scale this design to achieve as large a number as desired. Tononi argues that the reason such a camera is still not conscious, is because the information in the camera is not integrated – no information is passed between the photodiodes, and in fact the camera could store the same amount of information if the photodiodes where not physically connected at all.

Tononi's additional suggestion, then, is that if a physical system can store a high amount of information *and* if that information is integrated (in some sense to be further defined) then this may be the direct correlate of the informativeness and integration of conscious experience.

3 MEASURES OF Φ

Tononi has proposed two major measures of integrated information [2, 6]. It should perhaps be emphasized at the outset that neither integrated information, nor effective information (which Tononi also defines) are standard information theoretic measures. Instead, both the concepts and the measures are ones which Tononi and collaborators define, in order to capture their intuitions about the nature of consciousness. We emphasize this, as it might otherwise be supposed that integrated information and effective information are both well-known and well-defined physical quantities, and that all that is in question is whether these quantities relate to consciousness.

3.1 Φ Measure 1 (Φ_1)

Tononi's first measure of Φ is a measure of a static property of a neural system[3]. If Tononi is right, it would measure something like the potential for consciousness in a system. It cannot be a measure of the *current* conscious level of the system, for it is a fixed value for a fixed neural architecture, regardless of the system's current firing rates (e.g. in response to inputs or internal dynamics).

Tononi's first measure works by considering all the various bi-partitions (splits into two parts) of a neural system:

> "the capacity to integrate information is called *Φ*, and is given by the minimum amount of effective information that can be exchanged across a bipartition of a subset" [6]

That is to say, Tononi's approach requires examining every subset of the system under consideration. And then, for each subset, every bi-partition (split into two non-overlapping parts) is considered. Given a subset, S, and a bipartition into A and B, Tononi defines a measure called effective information (EI). Effective information uses the standard information theoretic measure of mutual information[4]. But rather than the standard mutual information measure which quantifies the information gain from taking account of the connectedness between A and B, Tononi's EI is a measure of the information gain which would accrue, if one considered the interactions between B and a *different* system, call it A', which is connected to B in the way in which A is, but whose outputs vary randomly across all possible values. The aim is to incorporate some sense of causality:

> "Since A is substituted by independent noise sources, the entropy that B shares with A is due to causal effects of A on B."

The logic of this sentence is perhaps not entirely clear[5], but the general idea is that the effective information from A to B shows the ability of A to affect B. Similarly, the EI from B to A shows the ability of B to affect A. The sum of these two is further defined as the effective information across the bipartition.

Now we can start hunting for Φ. First of all, for a given S, we look for the bipartition with the minimum (normalised[6]) EI. Then we define Φ(S) as the EI of that minimum information bipartition.

But Φ at this point is not yet true integrated information, in Tononi's sense. Next we must look for *complexes* – subparts which are not fully contained in any regions of yet higher Φ. According to Tononi, only complexes genuinely integrate information; Φ is a measure of how much information they integrate, and the Φ value of the *main complex* (the complex of highest Φ in the whole neural system) is the correct value to use for the integrated information of the system as a whole.

3.2 Commentary on Φ_1

The key points to note for now are the following. Φ_1 involves the definition of two novel informational concepts (effective information and Φ itself). Neither of these have anything like the range of applicability of standard concepts like mutual information or Shannon entropy (for they are defined in very specific ways, for a very specific system). On the other hand, Φ_1

[3] As with both Tononi's Φ measures, it is well defined only for a rather limited class of well-behaved systems; showing that it can be applied more generally would require further work.

[4] MI(A:B) = H(A) + H(B) – H(AB), where H(…) is entropy, a mathematically well defined measure of uncertainty. A decrease in entropy amounts to a gain in information (i.e. a decrease in uncertainty). MI in particular measures the information gain obtained from considering the interactions between A and B, as opposed to ignoring them. If there are no interactions between A and B (if they are independent systems), then the mutual information will be zero, otherwise it will be positive.

[5] After all, what we're really measuring is the mutual information (which is a symmetric measure, MI(A:B) = MI(B:A)) between B and a different system, A'.

[6] This is an attempt to avoid certain bipartitions being favoured for purely mathematical reasons. But see fn. 9 for more on the problems this process introduces.

certainly is a measure of information – this follows directly from the fact that it is built up from standard information measures such as mutual information. But the flip side of this is that Φ_1 has a perfectly good informational interpretation which follows from its definition. It is the reduction in uncertainty which an external observer would gain, if they took account of the interactions between A' (the perturbed version of A) and B, as opposed to treating these as separate systems (and vice versa for B' plus A). Since Φ_1 already has this meaning, it is unclear whether we can give it the additional meaning, as the system's own information, which Tononi wishes to. We will discuss this further below.

3.3 Φ Measure 2 (Φ_2)

In more recent work, Tononi and collaborators [2] have proposed a revised measure of Φ. This revised measure has some advantages over the previous measure, in that it can deal with a time varying system, providing a varying, moment to moment measure of Φ (which would correspond to a moment to moment measure of conscious level, if Tononi's approach works as intended)[7].

The revised measure of Φ is also defined in terms of *effective information*, though effective information is now defined quite differently from the version in the previous measure of Φ. In this case, effective information is defined by considering a system which evolves in discrete time steps, with a known causal architecture. Take the system at time t_1 and state x_1. Given the architecture of the system, only certain states could possibly lead to x_1. Tononi calls this set of states (with their associated probabilities) the *a posteriori* repertoire. Tononi also requires a measure of the possible states of the system (and their probabilities), in that situation where we do not know the state at time t_1. This is called the *a priori* repertoire. The *a priori* repertoire is calculated by treating the system as if we knew nothing at all about its causal architecture, in which case we must treat all possible activation values of each neuron as equally probable[8]. The *a priori* and *a posteriori* repertoires will each have a corresponding entropy value (for instance, if the *a priori* repertoire consists of four equally probable states, and the *a posteriori* repertoire has two equally probable states, then the entropy values will be two bits and one bit, respectively). This means that, in finding out that the state of the system is x_1 at time t_1, we gain information about the state of the system one time step earlier.

Tononi argues that this is a measure of how much information the system 'generates' in moving into state x_1. Having defined this measure of how much information the system generates, Tononi once again requires a measure of how 'integrated' this information is.

Therefore, he next observes that it is possible to arbitrarily decompose the system into parts. For each part (considered separately) a given current state can only have come from certain possible parent states. Similarly, for the system as a whole, the current state can only have come from certain possible parent states. Therefore we can ask, is there any possible decomposition into parts, such that the information from the system as a whole is no greater than the information from the parts separately? If there is, then we have found a way to decompose the system into totally independent parts.

In the case where the system does *not* decompose into totally independent parts, we can once again look for the decomposition which gives the *lowest* additional information from the whole as opposed to the parts[9]. Tononi calls this the *minimum information partition*. The effective information (the additional information given by the whole, as opposed to the parts) for the minimum information partition is then the Φ value for the system.

Finally, we can do an exhaustive search across all subsystems and all partitions[10], and once again we can define *complexes*. A complex is a system with a given Φ value, which is not contained within any larger system of higher Φ. Similarly, the main complex is the complex with highest Φ in the whole system – and the true measure of Φ (or consciousness) for the system is the Φ of the main complex.

3.4 Problems with Φ_2

In examining Φ_2, we note that many of the problems with Φ_1 still apply. Firstly, EI and Φ itself are defined in ways which are closely tied to the particular type of system being examined. Although Φ and EI are intended as general purpose concepts, the current mathematics has nothing like the broad range of applicability of standard information theoretic measures.

As before, Φ_2 is indeed a measure of information. But this follows from the fact that the procedure for calculating Φ involves mutual information, which is itself a well-defined information-theoretic measure. Where the Φ_1 measures the amount of information which an external observer could gain about one part of the brain, from another part, Φ_2 measures the amount of information which an external observer could gain about the earlier state of the brain, from the later state.

It is true that, by including a procedure for identifying the minimum information partition, Φ does give some indication of how functionally integrated the system is. But Tononi wants considerably more. He suggests that Φ is "information from the perspective of the complex itself" (p.17), and that it is information "that the system generates" (p.3), "independent [of the point of view] of any external observers" (p.3) [2]. Elsewhere, he goes as far as to claim that Φ "exists as a fundamental quantity – as fundamental as mass, charge, or energy" [1].

He also suggests that:

> "The intrinsic nature of integrated information, which only exists to the extent that it makes a difference from the perspective of the complex itself, is usefully contrasted with the traditional, observer-dependent definition of

[7] It has disadvantages too – including apparently allowing the (presumably continuous) stream of consciousness of a given system to reside in quite different parts of the system from moment to moment.

[8] In fact, this is not a true measure of our prior knowledge about the state of the system: a given causal architecture may make certain firing patterns simply impossible, in the normal time evolution of the system, whatever the inputs. Even if Tononi's EI were modified to take this into account, however, it would not address the objections to Tononi's interpretation of Φ given below.

[9] Once again, a normalisation factor is introduced. Otherwise asymmetric partitions will be disfavoured, and partitions into multiple parts will be favoured, for purely mathematical reasons. Unfortunately, as Barrett and Seth [7] point out, this normalisation itself introduces undesirable properties into the definition of Φ, and make it implausible that Φ as it stands really corresponds to any fundamental property of the world.

[10] At least in principle; in practice, this may well be far from feasible for neural systems of the scale of a real human brain.

information, in which a set of signals are transmitted from a source to a receiver across a channel (or stored in a medium), and their "integration" is left to an external human interpreter." [2]

Is it really true that Tononi has found a way to achieve point-of-view free information? We will suggest below that this can't be achieved. We also note that both measures of Φ are effectively self-information in the brain – the information is not necessarily about the world, at all. But there are good reasons to think that an agent's own information should be about the world.

We will examine these issues from two perspectives, below. Firstly, we will examine the well-known (though controversial) Bayesian interpretation of probability theory, and will argue that Tononi's measure cannot have the interpretation he wishes, if the Bayesian approach is correct. We will also note that this approach implies that an organism's own information is fundamentally about the world.

Next we will contrast Tononi's Φ with a more internalist approach to information. But even here, we will see that there are good reasons for thinking that Tononi's Φ is far from the whole story about consciousness, precisely because his measures are concerned only with interactions within the brain, and not with interactions between brain, body and world.

4 AN ENACTIVIST PERSPECTIVE ON INFORMATION

4.1 Probabilities are Subjective - Cox's Theorem

Jaynes [8] following Cox [9] (and earlier writers, including Keynes [10]) has presented strong arguments to show that the standard calculus of probability is actually the correct calculus for describing consistent reasoning in the face of subjective uncertainty.

Specifically, if we want real numbered values to represent a subject's credence in given propositions, and we wish the subject's reasoning to remain consistent with certain very basic common sense requirements, then it can be proven mathematically that the numbers which the subject uses must combine and interrelate according to the standard sum and product rules of probability theory:

$$p(A|B) + p(\neg A|B) = 1$$

$$p(AB|C) = p(A|C)p(B|AC) = p(B|C)p(A|BC)$$

A key point made by Jaynes, and Cox, is that probability theory under this Bayesian interpretation of the meanings of the symbols is actually more broadly applicable than probability theory under a frequentist interpretation. All of the frequentist applications of probability theory can be derived as special cases of the Bayesian theory; but the Bayesian theory remains consistent and applicable in many cases where frequentist theory says probabilities cannot be used.

The 'argument' between these two interpretations is not just a philosophical one, for the Maximum Entropy approach to statistics (which can be justified directly on Bayesian grounds, but cannot be justified at all within the frequentist approach) now has many highly successful applications in the applied physical sciences (in image processing, signal detection, and so on) [11].

4.2 Entropy is subjective

Given a complete and mutually exclusive set of possible outcomes, i, and probabilities p_i for each outcome in i, then the formula for the entropy H of this probability distribution is:

$$H = -\Sigma\, p_i \log (p_i)$$

This formula also has a clear interpretation, in terms of the amount of uncertainty represented by a probability distribution. We can see by inspection that the measure has the right broad properties: more options result in more uncertainty, and a more even distribution of probabilities also equates to more uncertainty. But in fact Jaynes [8] (following Shannon [12]) shows the measure is not arbitrary – simple logic, combined with careful mathematics, shows that it is the only reasonable and consistent mathematical measure of uncertainty, under some very minimal requirements for such a measure.

Note that nothing here has stepped outside the realms of subjectivist probability theory; that is to say, entropy is defined in terms of probabilities, and is well-defined when (and only when) probabilities are well-defined. So our interpretation of entropy will depend on our interpretation of probability.

To avoid being misunderstood at this point, the claim that entropy is subjective should be clarified. As Jaynes puts it :

"[Entropy] is "subjective" in the sense that it … measures uncertainty; but it is completely "objective" in the sense that it depends only on the *given data of the problem*, and not on anybody's personality or wishes." [8]

That is, given a certain partial state of knowledge, there is only one correct and consistent measure of one's uncertainty – the (maximised[11]) entropy.

4.3 Information is subjective

From this, it also follows that all the information measures Tononi builds on (and, indeed, all standard information measures) are also subjective, in the same sense. They are all defined as comparisons between probability distributions (the most simple information measure being just the arithmetical difference between 'before' and 'after' entropy values[12]).

Since information is fundamentally defined in terms of probability distributions, and since probability distributions fundamentally measure uncertainty from a given partial point of view, the Cox/Bayes view entails that states of the world do not ever intrinsically carry information – they only carry information from certain (partial) points of view[13].

As emphasized above, this does not mean that information becomes a matter of opinion. Once I've clearly defined my state of partial knowledge about a system (e.g. that any one of four distinct symbols may be transmitted next, and I have no reason

[11] Maximisation of entropy won't be discussed further here; but broadly speaking, it is the best (most self-consistent) approach for assigning initial probability values (something which frequentist probability theory is ill-equipped to deal with), when these would otherwise be underdefined by the data of the problem.

[12] Relative entropy, or Kullback-Leibler divergence, is arguably a more fundamental measure of information gain. It is defined in a more complex (but closely related) way, but it is still fundamentally a comparison between 'before' and 'after' probability distributions.

[13] A lot of the time, when working with information measures, we are therefore specifying how much information an idealised subject would gain, if they were in a specified state of uncertainty, and then gained a specified new piece of evidence (e.g. that symbol *x* arrived).

to think one more likely than the others) then there is an objective fact of the matter about the information available to me, in gaining new evidence about the system (e.g. the amount of information transmitted for any given symbol is two bits).

4.4 Information presupposes an integrated subject

Another key factor of the above analysis is that information theory *presupposes* the existence of a *coherently acting rational subject*, for it presupposes that we are dealing with an agent with the ability to understand propositions (A, B, C, etc. in P(A|B), etc.) and see when they apply to the world.

This point can be seen clearly, when we recall what Jaynes and others have noted [8, 10]: that probability theory is an extension of classical (Aristotelian) logic. Aristotelian logic formalizes the patterns of correct deductive reasoning (e.g. if A then B; A; therefore B); but it doesn't tell us what it is to understand a proposition and to apply it to the world in the first place. Equally, the logic of probability theory formalizes the correct patterns for both deductive (certain) and inductive (probabilistic) reasoning – but once again, the theoretical framework presupposes the existence of agents able to understand propositions and to perceive their applicability in the world.

The rational coherence *presupposed* here looks very like the integration which Tononi wants to explain (the second of his two unargued intuitions about consciousness). A single subject must be able to perceive, and understand the relevance of, multiple distinctions at once ('red', 'blue', 'chair', 'table', etc., etc.).

4.5 Where is information for a subject?

Less we be misunderstood, a further clarification is in order. It is often supposed that information for a subject 'must' be somewhere in the subject's brain. On the account of information proposed here, information, in the first instance, is something available to a rationally *behaving* subject. If we see a subject updating their credences rationally in the face of new evidence[14], and then acting rationally on their subjective credences[15], then we can apply to formalism of information theory to quantify how much information the subject gains (or would gain), in a given situation.

It is at least arguable, then, that "information for a subject" is a different (and more fundamental) concept than "information in a subject's brain".

However a traditional, and still widespread, view in cognitive science supposes that information in brain states (the information which a third party observer can find out, about the world, from brain states) *is* the information for the subject (the information which a subject has, about the world). This is the essence of representational theory of mind in cognitive science. The argument here is not yet resolved. For instance, the experiments of Beer [14] and Izquierdo and Di Paolo [15] seem to suggest strongly that information for the agent (i.e. what the agent knows about, as manifest in its actions) need not be present as information in the brain (i.e. what an informed third party observer can find out about the state of the world, by examining the state of the brain). In one example [15], a simple agent makes a decision as to whether to catch or avoid a certain falling shape. This decision becomes 'locked-in' at a certain point during each catch/avoid trial. But we are guaranteed that an external observer *cannot* work out which decision has been made, just by looking at the 'brain', for the neural architecture has no persisting internal state to represent its decision. This agent 'makes a decision' by actually moving to a different place in the world, i.e. by making use of the external dynamics of the task.

Examples such as these tend to support the claim that there really are two levels of analysis: information for the agent, and information in the agent's brain – and that the two need not coincide, in real-world tasks.

However, many would still argue that truly complex cognitive tasks are "representationally hungry" [16]; i.e. they are tasks where the information which the agent possesses about the world must be represented in the agent's brain (i.e. decodable, in principle, just from the brain, by an external observer, even though the decoding may be far from trivial).

In the next part of this paper, we look at another view on consciousness; one which presupposes (as does Tononi) that information for a cognitive agent can be found as information in the agent's brain. (Therefore assuming that the two levels of analysis argued for above don't come apart in real cognitive agents.)

Interestingly, even within this more standard, 'internalist', framework, we find that there are still reasons to think that Tononi's view of consciousness is incomplete, because it ignores interactions with the world.

5 THE INTERNALIST PERSPECTIVE

The internalist perspective taken here relates to 'synthetic phenomenology' work published elsewhere [17]. This has previously been discussed as 'an axiomatic theory of consciousness' [18] in which internal states that may be 'used' by the organism in its interaction with the world and give the organism a point of view of being in an 'out-there' world. An example of a usable internal state is one which depicts the tail of a perceived snake which causes the eye fovea subsequently to move to the head of the snake to determine whether the perceiving organism should run or stay. For the purposes of this section of the paper, this is what we mean when we say that the depictive state is phenomenal: it has information about the world which may be necessary to cause rational behaviour in the world[16]. In this section of the paper we consider a structure with, at best, very simple 'rationality', but which has the property of creating internal states which iconically represent external events through learning. We assume that the world is an automaton which presents its states in time and through a limited bandwidth interface to the learning organism. Part of the 'meaning' of this world is that there is a structure that links its states. At any moment, the task for the learning organism is to generate information by identifying not only the state among all states it

[14] As noted at the beginning of the previous subsection, the ability to take in evidence is something presupposed in the formulation of probability theory.

[15] To interpret behaviour as rational requires that we additionally postulate some cost/utility function – i.e. we incorporate aspects of decision theory. It is true that there are right (rational) things to do, *given* a utility function. But there is no right answer as to which utility function an agent should use. So interpretations in terms of rationality are always to be evaluated in terms of usefulness (relative to other predictive strategies) [13] and range of applicability.

[16] We argue for this usage of the word *phenomenal* in [17].

has experienced at the interface (a facet of IIT) but the linking state structure to which it belongs. For example, consciousness of the front of a car on the road generates some static information, but if the next state is of the car is bigger, the event is identified as part of a 'danger' state structure of the car getting closer, while if the car gets smaller, this is part of a structure with a meaning of 'safe'. Therefore here we broaden the concept of information integration as being between an organism and the environment in which it is embedded. This allows us to intuit that there are levels of the organism being informationally integrated with the world at the time of learning which are usable and levels where the integration fails either to have usable states or to internalise the structure between these states.

To best illustrate this we define a neural phenomenal automaton $\mathcal{P}$ which 'observes' the world and which can be defined classically as a 5-tuple:

$\mathcal{P}$: <I, Q, Z, δ, ω >

Where *I* is the set of all possible inputs on an *n*-bit interface:

$I = \{i_1, i_2, ... i_{/I/}\}$, where /*X*/ is the magnitude of set *X* which for /*I*/ is 2^n ,

$Q = \{q_1, q_2, ... q_{/Q/}\}$ is the set of all possible inner states,

$Z = \{z_1, z_2, ... z_{/Z/}\}$ a set of possible outputs.

δ is the mapping (*I* x *Q)* into the 'next' value of *Q*,

ω is the mapping of *Q* into *Z*.

For the states of the system to become phenomenal, we assume that the state variables are weightless neurons [19] and that 'iconic' training takes place as described next.

Given a weightless system assume an *n*-bit input and $i_t \in I$ a pattern that appears at that input at time *t*. Say that the network is in state $q_{t-1} \in Q$. Iconic training is the forcing of $i_t, q_{t-1} \to q_t (= i_t)$. This effectively transfers i_t into the state structure of the network predicated on the net being in q_{t-1} and the input being in i_t. We note that iconic training causes *I* to become a subset of *Q*. In a weightless net, generalisation takes place in the sense that a pair $(i_a, q_a) \to i_j$ where $(i_j, q_j) \to i_j$ is the training pair which best matches (i_a, q_a) (usually in a bit-for-bit way).We define a *phenomenal system* as one in which $I \subset Q$ forms a *closed* state structure with δ that only generates states within *M*.

But it is not sufficient that the state structure of $I \subset Q$ be just closed. To be phenomenal it must be *about* the world as seen at *I* in terms of mimicking the sequential machine as seen from *I*. An example might help at this stage.

Example

I is established as a square binary window of 40x40 bits. In this example the world presents a state structure shown in fig. 1.

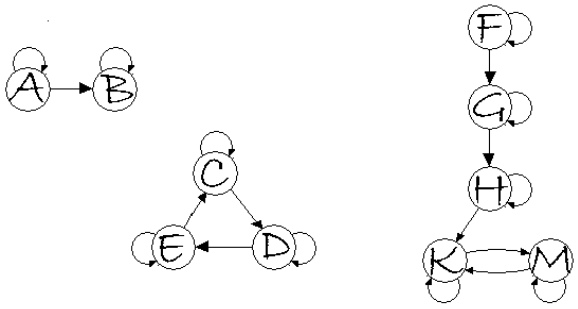

Figure 1: Structures of world states. The letters are 40x40 bit images at *I* and are not symbols. All states can 'linger' as well as transit to another state. So, looking at the top left behaviour, the world can sustain A from which it can change to B and rest there, but not return to A. Four behaviours of the world are shown.

The iconic training of the weightless network proceeds as follows.

To train on a re-entrant state *x* present at *I* (being sustained in time) whatever the state on *Q*, say, *q*, an iconic transfer is applied which means that the general learning step

$(i_j, q_j) \to i_j$ becomes $(x, x) \to x$.

Then when the input changes to state *y* the iconic training step is $(y, x) \to y$.

We now show by experimentation (figure 2) that connectedness parameters in the net not only bring about a loss of uniqueness and indivisibility in static states [3], but disrupt the ability of the net to identify the state structure of the world.

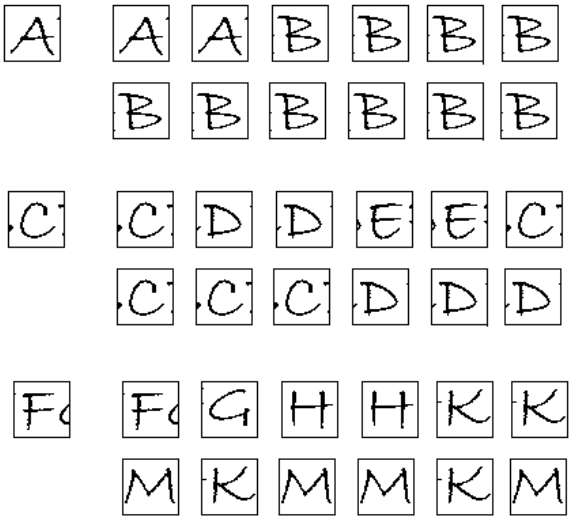

Figure 2: The inner response of a fully connected net. The left image is at the input for one time step only, after which it is replaced by noise to disconnect the automaton from the world. This noise then randomises the transition back into the current state or on to the next state.

The network parameters in figure 2 are set to the maximum effectiveness of every neuron being connected to the outputs of all the others in the state and the inputs of the automaton before learning. This ensures that the iconic learning cannot lead to

contradictions.[17] So it is seen that exposure to the state structure of the world is 'understood' by the learning network which identifies the four distinct behaviours of the world and generates internal representations that fully represent these behaviours. It could be said that the organism generates information by properly integrating with the world through an internal activity that is 'about' the world and is therefore phenomenal. This mechanism fails as the connectedness is reduced in the learning automaton and between the automaton and the world. The resulting state sequences are shown in Fig 3.

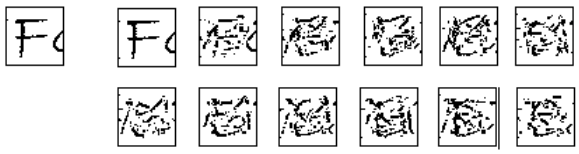

Figure 3. Now each neuron samples the output of 25 neurons (5 x 5 from the state array) and from a corresponding 5x5 array from *I*. The response to F in the third state group of fig.1 is shown and should be contrasted with the third result in fig.2. Each neuron samples the output of 25 neurons (5 x 5 from the state array) and from a corresponding 5x5 array from *I*.

Although we have not yet produced quantitative analyses of these results, it can be clearly seen by eye that not only are the individual states not sustained, but the sequences of the world state structure are not properly recovered.

5.1 Observations on the Internalist Perspective

- In contrast to the stance by Balduzzi and Tononi [20] that qualia are captured by the geometry of the information integration between neurons during state changes, our perspective requires that the content of a state (qualia of sorts) result from information integration across the world/organism interface at a time that learning takes place.
- In other words, it can be said that we look to the internal state structure of the organism to be 'about' the state structures of the world in the sense that it has identified the dynamic structure of the states it observes and generates information by identifying one state structure among many.
- Using a sufficiently rich connectedness both within the neurons of the inner network and their connection to the external world, we have shown that how this internal representation happens. Lowered richness leads to failures.
- We maintain that integration and its failure loosely correspond to high and low effective interchanges of information as in IIT theory. Clearly there is a need to develop some predictive formulations here – a topic for future research.

[17] As an aside this has to be distinguished from a 'fully connected Hopfield network' as discussed by Balduzzi and Tononi [2]. In our work, while the network is physically fully connected before training, in training itself effectively a vast number of interconnections is rendered ineffective, so connectedness calculations include the effect of training. This is left for further work, here we concentrate on empirical results.

6 CONCLUSION

We now briefly summarize the responses to Tononi's Φ measure given in this paper. Firstly, we have noted that Φ (and the closely related Effective Information, or EI), are not (or certainly, not yet) general purpose information theoretic concepts. Their mathematical definitions are closely tied to the analysis of specific physical systems (and, moreover, these mathematical definitions have changed quite considerably, when different physical systems have been analysed). We accept that Tononi's measure is a potentially useful measure of how functionally integrated a system is. But we have questioned whether it really measures a fundamental quantity, corresponding to the system's own, conscious, information as claimed.

Adopting an 'externalist' perspective, we have noted that under one compelling understanding of information theory, Φ is the wrong type of measure to capture a subject's own information. On this Bayes/Cox perspective, the concept of the information 'in' a physical state is a concept of how much information a subject can gain from *examining* that state. This is to be contrasted with the concept of information for a subject, which is the concept of how much information a (rationally behaving, Bayesian) subject gains about the world when encountering certain evidence. This latter is the more fundamental concept, and it is defined in terms of how a rational subject acts (or would act, if appropriately tested – a subject can possess information without having to show it behaviourally). We have suggested that these two levels may not just be logically distinct, but actually distinct, in the case of brains and behaviour: a subject's information may involve body and world in ways which mean that the subject's information simply isn't decodable from the subject's brain state.

However, since there are several controversial steps in the above analysis, we also examine Tononi's Φ from a more 'internalist' perspective. Even on this perspective, we demonstrate empirically that the information which a subject can gain about the world depends not only on the level of integration in the subject's brain, but also on the level of integration between subject and world, across the sensory interface.

Therefore, in common between these two views are the claims that conscious information is always about the world (i.e. not just something internal), and that consciousness fundamentally involves interaction with the world. These are results would follow from the nature of information itself, if one accepts the more controversial 'externalist' arguments we have given, and are anyway demonstrated empirically, using a rather less controversial 'internalist' approach.

These results also demonstrate clearly that it is quite possible to be sympathetic to Tononi's intuitions about the nature of consciousness (as we are), without having to follow Tononi down the route of accepting that Φ, as formally defined, corresponds directly to consciousness itself.

Tononi's Φ *may* still turn out to be a good objective correlate of consciousness. But showing this would require that Φ be defined in a much more general purpose way than it has been to date; and even then, it might turn out that high Φ is an explanatory correlate [21] of something more fundamentally associated with consciousness (for instance, of the instantiation of consciousness understood axiomatically [22], or simply of the presence of coherent, complex rational behaviour [23]).

ACKNOWLEDGEMENTS

This work was supported by a grant from the Association for Information Technology Trust.

REFERENCES

[1] Tononi, G. (2008), Consciousness as Integrated Information: a Provisional Manifesto. *Biol. Bull*. **215**: p. 216-242.

[2] Balduzzi, D. and G. Tononi (2008), Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLoS Computational Biology*, **4**(6): p. 1-18.

[3] Aleksander, I. and D. Gamez (2009). Iconic Training and Effective Information: Evaluating Meaning in Discrete Neural Networks. in *Proc. AAAI Fall Convention - Brain Inspired Cognitive Systems Symposium*.

[4] Shear, J. and R. Jevning (1999), Pure Consciousness: Scientific Exploration of Meditation Techniques. *Journal of Consciousness Studies*, **6**(2-3): p. 189-209.

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

[6] Tononi, G. and O. Sporns (2003), Measuring Integrated Information. *BMC Neuroscience*, **4**(31).

[7] Barrett, A.B. and A.K. Seth (2011), Practical Measures of Integrated Information for Time-Series Data. *PLoS Comput Biol*. **7**(1).

[8] Jaynes, E.T. (2003), *Probability Theory: The Logic of Science*, ed. G.L. Bretthorst. Cambridge: CUP.

[9] Cox, R.T. (1946), Probability, frequency, and reasonable expectation. *Am. Jour. Phys.*, **14**: p. 1–13.

[10] Keynes, J.M. (1921), *A Treatise on Probability*, London: MacMillan.

[11] Erickson, G.J. and C.R. Smith (1988), eds. *Maximum-Entropy and Bayesian Methods in Science and Engineering*. Kluwer: Dordrecht.

[12] Shannon, C.E. (1948), A mathematical theory of communication. *Bell System Technical Journal*. **27**: p. 379-423 and 623-656

[13] Dennett, D.C. (1987), *The Intentional Stance*, Cambridge, MA: MIT Press.

[14] Beer, R.D. (2003), The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 2 **11**(4): p. 209–243.

[15] Izquierdo, E. and E. Di Paolo (2005), Is an Embodied System Ever Purely Reactive?, in *Proceedings of the 8th European Conference on Artificial Life*, M.S. Capcarrere, et al., Editors, Springer-Verlag: Berlin. p. 252-261.

[16] Clark, A. and J. Toribio (1994), Doing Without Representing*? Synthese*. **101**: p. 401-431.

[17] Aleksander, I. and H. Morton (2007), Depictive Architectures for Synthetic Phenomenology, in *Artificial Consciousness*, R. Manzotti, Editor, Imprint Academic: Exeter.

[18] Aleksander, I. (2007), Why Axiomatic Models of Being Conscious? *Journal of Consciousness Studies*. **14**(7): p. 15-27.

[19] Aleksander, I., *et al*. (2009) A brief introduction to Weightless Neural Systems. in *Proceedings of ESANN 2009*. Bruges.

[20] Balduzzi, D. and G. Tononi (2009), Qualia: The Geometry of Integrated Information. *PLoS Computational Biology*. **5**(8): p. 1-224.

[21] Seth, A.K. (2009), Explanatory correlates of consciousness: Theoretical and computational challenges. *Cognitive Computation* **1**(1): p. 50-63.

[22] Aleksander, I. and B. Dunmall (2003), Axioms and Tests for the Presence of Minimal Consciousness in Agents. *Journal of Consciousness Studies*, **10**(4-5): p. 7-18.

[23] Beaton, M. (2009), *An Analysis of Qualitative Feel as the Introspectible Subjective Aspect of a Space of Reasons* (D.Phil.: University of Sussex).

A Role for Consciousness in Action Selection

Joanna J. Bryson[1]

Abstract. This paper argues that conscious attention exists not so much for selecting an immediate action as for focusing learning of the action-selection mechanisms and predictive models on tasks and environmental contingencies likely to affect the conscious agent. It is perfectly possible to build this sort of system into machine intelligence, but it is not strictly necessary unless the intelligence needs to learn and is resource-bounded with respect to the rate of learning vs. the rate of relevant environmental change. Support of this theory is drawn from scientific research and AI simulations, and a few consequences are suggested with respect to self consciousness and ethical obligations to and for AI.

1 Introduction

Consciousness is first and foremost a culturally-evolved concept of uncertain age and origin (14). As such it is not at all clear that the many things we call *consciousness* are truly aspects of a single psychological phenomenon. Even were they to be so, we would not necessarily know the phylogenetic priority between the various traits we identify with consciousness.

For the purpose of this article at least, I will focus on a completely functionalist account of consciousness and intelligence more generally. Consciousness is one evolved element of intelligence, and presumably serves a role within the cause of intelligence. I will start from the assumption that the cause of intelligence, its essential role, is primarily to do the right thing at the right time. Intelligence survives natural selection entirely as a consequence of the advantage the actions it generates gives its host, and their outcomes in terms of the agent's (or at least, the agent's genes' (12, 37)) survival and ability to reproduce.

If consciousness is adaptive in nature then it could well be useful for AI as well. This might not be true if for example consciousness is essentially a mechanism for implementing serial processing on the massively-parallel architecture which is the vertebrate brain. Since AI to date has tended to be minimally concurrent we might even in that case need some kind of "reverse consciousness" to harness the power of concurrency with our sequential systems.

In this paper though I analyse a theory that consciousness is a strategy to combat the combinatorics of the search for appropriate actions available to agents capable of learning new strategies. I have previously argued that there exists a class of reaction time results that result not from the cognitive complexity of the task being performed, as is generally postulated. Rather delays in processing reflect an allocation of time by the learning-competent agent to on-line search for a better action (5, 6). The amount of time allocated to this search in real-time by an individual depends on its confidence with respect to the task. The more certain an animal is, the less time it allocates to searching for a better solution or prediction concerning the situation. There are also species-specific and life-history components to the duration of the search. An assumption which we have yet to demonstrate in the laboratory is that the period of search correlates to conscious attention to the task and the feeling of awareness.

If we are correct in our accounts, this feeling-of-awareness part of consciousness can be shown to be shared with monkeys, rats and presumably many other intelligent vertebrates, though they may spend less time in this state and more in a state of "automatically" generating behaviour than the average human. Further, to the extent that we are willing to call this consciousness, this addresses the question of the utility of machine consciousness as well. Where machines benefit from applying resource bottlenecks to searching for new solutions, they might also benefit from a similar strategy. This would make a machine also functionally aware of a strategically-limited subset of its environment, rendering it much like a conscious human.

In this paper I seek to clarify this theory and then examine its implications. In Section 2 I describe conscious attention and cognition in an evolutionary context. In Section 3 I explain the details of and evidence for the theory. In Section 4 I describe its application to machine intelligence, and in Section 5 I briefly examine the theory's implications for self consciousness and ethical obligations.

2 Functionalism, Evolution, Cognition and Learning

If consciousness is useful to intelligence and intelligence is useful to survival, then why are we not conscious of everything all the time? Many theories of consciousness assume that it requires some sort of expensive resource which must unfortunately be limited, perhaps by metabolic cost or by the size of heads during child birth. Consciousness therefore inherits this scarcity and must be preserved to be directed with care at only the most important problems.

In general, where we see a variety of solutions in biology this indicates a tradeoff between the costs and benefits of a trait, allowing the perpetuation of roughly equally-fit variation along the axis projected by this tradeoff. The best-known example of this is the tradeoff between the number of offspring an individual can have and the amount of care it can invest in each of them. Certainly the extent to which species rely on cognitive strategies for selecting appropriate actions is highly variable. Cognition — by which I mean any real-time, online modelling of the expected outcomes across some range of behaviour alternatives — is a broadly unpopular solution ignored by plants and single-cell organism, though both of these are capable of expressing behaviours in response to their environment. Bacteria move towards or away from substances and behave socially with other bacteria to improve their situation and prospects for preserving their genes, sometimes at the cost of self-sacrifice (37). Plants are capable of responding not only to light and nutrients but also to

[1] University of Bath, England, United Kingdom email: j.j.bryson@bath.ac.uk

pheromones of other (e.g. host) species of plants, and to direct their growth accordingly (36).

The tradeoff that follows from my proposal in the introduction is that cognitive strategies generally cost time – time for cognitive processing delays action. Time is expensive. A delay may mean that another agent takes advantage of a situation before you. Heubel et al. (22) demonstrate that mate competition may explain the failure of male mollies to learn to discriminate the Amazon molly *Poecilia formosa* even though 'mating' with these females gives them no fitness benefits. The time it takes to discriminate the Amazon mollies from females of the male's own species is more valuable than the cost of insemination, because those that hesitate are beaten to available conspecific females by those who do not. Even where there are no other competing agents, the situation may change before you are yourself able to take advantage of it. For example, a strategy for crossing roads must involve reaching decisions about recognising safe windows for crossing before that window disappears.

Psychometric research indicates that there is something intrinsically slow and also something noisy about biological consciousness (11, 27). If this is true, then even within a highly-cognitively-resourced organism it would still be adaptive to use conscious strategies only when other mechanisms fail. Norman and Shallice (27) describe essentially an interrupt-driven theory of consciousness where the special attention is only utilised in some circumstances, for example when a task is unfamiliar or particularly important to get right. The full version of their theory is at odds with the reports of skilled athletes, artists and musicians that their accuracy is *higher* when they are not attending to detail. However, humans and other cognitive species certainly do seem to turn our attention not only to tasks that are not familiar, but to any surprising stimulus. This phenomenon underlies the popular looking-time strategy for getting at what infants and other non-linguistic animals know (30, 34). Again, here we see the experimentally-validated premise that organisms attend longer to things that are unfamiliar, or — in machine learning terms — that they were unable to predict.

What then is the advantage of cognitive approaches that compensates for this loss of time? Apparently, plasticity — the ability to solve problems and take advantage of opportunities that change more rapidly than other ways of acquiring action selection rules, e.g. evolution or implicit learning, can manage.

3 Timing, Awareness and Learning

In the previous sections I have argued that a fundamental cost of consciousness is time. Assuming that consciousness is engaged in some form of computation, then the source of this time penalty is combinatorics (33). There are potentially-infinite combinations of contexts to consider as triggers for an uncountable set of nuanced actions. However, no agent computes all possible actions or explanations. Organisms are not only restricted by time. Evolution has given organisms restricted action and perception abilities, and it further restricts their capacities to learn to associate actions and perceptions even within their species' competence. As the behaviourists proved while failing to validate Skinner's behaviourism, even simple stimulus-response conditioning does not work for all stimuli to all responses. Pigeons can learn to peck for food, but cannot learn to peck to avoid a shock. They can, however, learn to flap their wings to avoid a shock, but not for food (23). Rats presented with 'bad' water learn different cues for its badness depending on the consequences of drinking it. If drinking leads to shocks, they condition to visual or auditory cues, but if drinking leads to poisoning they learn taste or smell cues (17). These limitations are not handicaps, but rather should be seen as a set of prior expectations that accelerate learning in most situations that animals of a species are likely to find themselves in.

The amount of time allocated to cognition is set by at least four different factors. First, as I proposed in the Introduction and as is suggested by reaction-time performance on some specialised tasks, individuals may allocate more attention for longer when they are less certain that they know how to behave in a context. Second, as implied my account in Section 2, the emphasis placed on cognition by a species as a whole is a part of its adaptive suite (26, 35). Hauser (21) argues that species of primates such as tamarins that chase fast prey like insects have limited learning potential because they have evolved to be disinhibited — to minimise response time at the cost of a capacity to learn. This suggestion is also supported by Bussey et al. (10) who report that rats can only be trained to do task learning using a touch screen if an obstacle is placed in front of the screen. Being slowed down to crawl over the obstacle apparently gives them time and / or attention — the mental presences — to be able to notice a reward schedule.

A similar failure to notice reward schedules triggered my own theory of conscious attention. This time, the failure to learn is in elderly macaque monkeys. Rapp et al. (29) show that aged rhesus macaques have two peculiarities in their task-learning performance. First, they do not exhibit a reaction-time effect traditionally attributed to computation the task requires, yet their performance is identical to younger animals that do show this effect. Second, the aged macaques do not learn new behaviour when their reward schedule changes, unlike the younger animals that show the delay.

The task concerned is transitive inference (TI). This is a standard cognitive task introduced to developmental psychology by Piaget (28) and to experimental psychology through Bryant and Trabasso (3). TI formally refers to the process of reasoning whereby one infers that if, for some quality, $A > B$ and $B > C$, then $A > C$. Piaget described TI as an example of concrete operational thought, but Trabasso demonstrated it in pre-concrete-operational children. It has now been demonstrated in a variety of animals as well as young children (18). Performance of this "pre-cognitive" version of TI has a number of associated characteristics. The one most relevant to the present discussion is the Symbolic Distance Effect (SDE). The SDE is a reaction time (RT) effect. When subjects execute a transitive comparison, they operate *faster* the further away two items are in the implied sequence. For example, a correct decision on BD would be slower than one on BE, even if E is not the last item in the sequence[2] If TI were performed by simple inference, then items further apart would be expected to take *longer*, because more inferences have to be performed. That they are in fact faster helped motivate theories that transitivity learning is somehow innately sequential. Researchers have hypothesised that the subjects somehow recognise the sequential organisation of the stimuli and represent it internally in such a way that further-removed stimuli were easier to discriminate (3, 39).

However, the SDE is not a reliable individual effect, only an aggregate one (25). This already throws doubt on any computational account of the SDE. Bryson and Leong (9) demonstrates that a stimulus-action model proposed originally by Harris and McGonigle (19) can better account for the difficulties subjects have learning the initial stimuli pairs in the first place. It is actually fantastically difficult for cognitively-limited subjects to learn that a single stimulus is good in some situations and bad in others. Getting a substantial

[2] End items are by far the easiest stimuli in TI, because unlike intervening items they are uniformly rewarded. Thus TI studies generally exclude end items from study.

number of individuals to pass criteria on learning the pairs requires a careful learning regime. Bryson (6) shows that if we assume that animals hesitate before acting on their training in proportion to their certainty about which stimulus should take precedence, then the SDE can be replicated in aggregate (and not in individual, just as in live subjects) with this rule-associative model.

Why then do the elderly monkeys used by Rapp et al. (29) show neither SDE nor learning when a reward schedule has changed? I speculate that as monkeys advance in age, the probability that they have learned tasks well increases so the probability they will benefit from inhibiting acting decreases. Their very survival to an advanced age effectively increases their certainty in their actions, though the operation is neurological reduction of capacity for inhibition rather than cognitive certainty. This comes at a cost of reducing their capacity for learning in unexpected settings.

How does this relate to consciousness? Until we can replicate the no-SDE results in humans, we cannot be sure. But given both the monkey and the rat results it seems intuitive that the lack of SDE correlates with the lack of conscious attention. Few would argue against the claim that consciousness plays an intrinsic role in some forms of learning. Yet implicit learning can evidently take place and people can act in response to things they learn without having an explicit model of what they are doing. Some researchers report detectable differences in the quality or reapplicability of what is learned implicitly (1, 24), but at least to a superficial level the differences are often indistinguishable in the context of the task learned itself (32). What I am claiming here is that there exists a class of learning tasks that are only likely to be achieved when conducted with conscious attention. This class includes at a minimum the capacity to detect better strategies even during the performance of familiar tasks. This learning takes time, and this time is allocated by the individual in proportion to their certainty about the performance of the task. This is the third factor in the allocation of time for cognition mentioned at the beginning of this section.

The final, fourth factor is similar, but one we are more aware of and find less surprising. When we are aware there is a need for a rapid decision, we can make one. When we do so, we are also more likely to make errors (2, 31). Again, in humans this is a conscious as well as a cognitive phenomenon, but not one I will touch on further in this article.

4 How Much Machine Consciousness Does AI Need?

As I promised in the introduction, this paper is not about every aspect of consciousness. One of the advantages of AI and simulations more generally is that we can decompose evolved entities into their constituent parts, then attempt to demonstrate their resynthesis. If the resynthesis produces comparable results, we have a viable hypothesis. If our model is the simplest one that accurately describes the natural phenomenon it models, then it should be taken seriously.

The previous sections argue that conscious awareness — presence in the moment — such as is linked to the formation of episodic memory is correlated with the ability to learn not only episodes but also new reward schemes for task learning. Dennett has called consciousness a spotlight; my theory shifts the metaphor slightly to that of a searchlight. Action selection would in many cases go forward in the same way without the searchlight, except that it would in fact be *faster* in the darkness. The process of search requires not only special cognitive capacities but also time.

From a computational or machine learning perspective the advantages of this kind of system is easy to justify. Suppose we have a system which learns, but it cannot learn fast enough to build a complete model of its environment. This might be either because its environment keeps changing, or its life is short and its environment is complex, or because its rate of action depends on the complexity of its model so it needs to keep its model simple by constantly generalising it and forgetting something of the past. At any rate, the system needs to choose a subset of its environment to concentrate its learning ability — its learning *attention* — on. What would be a good set of criteria? Two obvious ones would be:

1. It should focus attention on the actions it is currently taking. This makes sense because any action it takes now it is likely to need to take again in the future — the things that it is acting upon are quite likely to be of some significance to it.
2. It should focus attention longer on things that it attends to but cannot predict.

If we combine these rules with the natural predispositions we find in nature to focus attention at least briefly on unexpected, loud or novel sounds or visual motion, then we might get quite an effective model of animals like grazing deer or cows. If we added in a drive to actively explore the manipulation of novel situations and affordances, we could simulate more creative species like predators or primates.

Of course a pressing concern from an AI perspective is — where in the action-selection process should the inhibition happen? The answer might seem to be obviously somewhere towards the beginning, since if a new perspective or alternative *is* discovered in the time allocated, selection can be improved. However note that in real animals and children, "looking" knowledge is not perfectly correlated with acting knowledge (30), and indeed some kinds of learning experiences do not seem to affect action selection until after a night's sleep (16). If neuroscience research like Shadlen's is representative of more complex tasks, then it really may be simply a general and ubiquitous slowing of the action selection process, and the advantages of insight may just be happenstance where they occur in time. It seems to me more likely that a candidate action is chosen quickly and then its execution is inhibited while the perceptual cues that elicited that response and the expectations driven by the intended action are allowed to play themselves out in the agent's working memory to see if alternative strategies become more attractive or alternative explanations seem more likely. If a better resolution does emerge the agent might be described as experiencing insight as it flushes its old plan and selects a new one.

5 Implications: Self Knowledge, Language and Ethics

Obviously there are many other aspects to the public concept of *consciousness* than these periods of awareness and basic capacities for learning models and correlations. I would now briefly like to talk about how some of these may follow from what I propose to be the most basic aspect of conscious attention.

The most obvious self consciousness isn't just consciousness, it's consciousness of the self, something that obviously requires a capacity for consciousness *and* a concept of self. In our culture, acquisition of the self concept is of course facilitated by language and shaped by culture. I stand in complete agreement with the recent work of Dennett (15) and more generally with the Extended Mind Hypothesis discussed by Wheeler (38) that consciousness and cognition more broadly are significantly enhanced, extended by and dependent on

material and social culture. But I do not think that this essential aspect of consciousness attention requires language or culture. Further, I doubt that consciousness is necessary for AI to exploit language and culture where those are able to be learned by brute force rather than in a systematic, task-driven way. I would argue that Google Search is absolutely an AI application that exploits human culture, but I don't see a reason to refer to Google as conscious.

To return to self consciousness, I doubt also, given the difficulty that children and even adults have in learning that every person *is* a person just like they are, that species without human language or culture do reliably achieve self awareness. Some individuals of social species do seem to show self consciousness, but I wouldn't take that as indicative that every individual is able to apply the rules it has learned to reason about others' behaviour to reasoning about its own. Google on the other hand has many searchable representations of itself and treats itself exactly like any other company or web presence.

One impediment to relatively simple explanations of attention and the concept of self such as those above is that our culture has an enormous amount of moral and ethical associations linked with consciousness. It is easy to imagine why there would be a confounding of consciousness with ethical obligation. Ethics is an evolved mechanism for sustaining societies, and it is most efficient when it appropriately allocates responsibility. Those who are aware are more likely to be responsible than those who are not, and also are more likely to be affected by our actions towards them. Most of our actions such as speech and gesture have relatively little impact on someone not aware of them. Only the conscious can be moral agents, but that does not necessarily imply that *all* conscious entities must be treated as moral agents.

Similarly, the technical definition of suffering involves the requirement that an animal's behaviour changes for the worse even after the disphoric situation (20). Clearly by the definitions given above this could only happen if the agent was learning (or attempting to learn) new behaviour while in the unfortunate situation. Thus this sort of conscious attention is necessary for an agent to experience suffering. But again, it is not sufficient. Even humans in particular neurological states do not suffer when they experience even severe pain (13). It is hard to comprehend some of the effects of anaesthetics, but easier to imagine building a machine that could be able to learn to perform tasks more generally but not to suffer.

In fact, my own opinion is that we are obliged when we make intelligent machines to make ones we are not obliged to (4, 7, 8). We can avoid uniqueness of body, and where there is uniqueness of mind we can ensure it is backed up appropriately. Further, any machine we build we will have built, and even if it acquires new goals we will have determined the means by which it acquires them. In this, machines and artifacts more generally are fundamentally different from the agents that evolved naturally along with us, including other people. In my opinion we should always view ourselves as essentially responsible for machines. Unlike the ordinary human process of children aging and becoming responsible first for themselves, then for their parents, I see no reason to replicate this process with AI. As I said, ethical systems have co-evolved with our societies. Now as our societies change rapidly, much of this 'evolution' is through deliberated legislation. I believe the most stable solution for human society is to value humanity over robots and maintain our responsibility for the machines we make. Otherwise there will be a moral hazard for people to commit violence and vandalism through their machines. Whether the machines are capable of learning while they are acting has little impact on the consequences for human society if we allow each other to displace our responsibility onto our creations.

6 Conclusion

In this paper I have argued that the most essential part of what we ordinarily call *consciousness* — that part that generates awareness of the moment and episodic memory — is a learning system associated with but not necessary for action selection in mammals. It provides a capacity for learning subtle contingencies in action selection — for noticing (for example) that a reward schedule has changed within an apparently-familiar task. I have suggested that the reason we are not conscious of everything at all times is simple combinatorial complexity — the fact that learning takes time and time is valuable.

I have suggested that machines will need this sort of attention only to the extent that they need to learn new skills or models and are limited in their ability to learn so need a heuristic for focusing their available capacity. In that case I suggest that the heuristic that has evolved for us is likely to be useful for them as well — to allocate attention on the actions you actually perform, and in time in proportion to your uncertainty about your next action, or to put it in another and more generally useful way, to predict changes in your immediate environment, including those expected to result from your action.

I have finally argued that this sort of attention is necessary but not sufficient for a variety of other phenomena we associate with consciousness — particularly ethical phenomena. It is however neither necessary nor sufficient for the concept of self in AI, but almost certainly precedes it in human and animal cognition.

ACKNOWLEDGEMENTS

I would like to thank the referees for their comments which helped improve this paper.

REFERENCES

[1] Diego Alonso, Luis J. Fuentes, and Bernhard Hommel. Unconscious symmetrical inferences: A role of consciousness in event integration. *Consciousness and Cognition*, 15(2):386–396, June 2006.

[2] Rafal Bogacz, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathon D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4):700, 2006.

[3] Peter E. Bryant and Thomas Trabasso. Transitive inferences and memory in young children. *Nature*, 232:456–458, August 13 1971.

[4] Joanna J. Bryson. A proposal for the Humanoid Agent-builders League (HAL). In John Barnden, editor, *AISB'00 Symposium on Artificial Intelligence, Ethics and (Quasi-)Human Rights*, pages 1–6, 2000.

[5] Joanna J. Bryson. Crude, cheesy, second-rate consciousness. In Mark Bishop, editor, *The Second AISB Symposium Computing and Philosophy*, pages 10–15, Edinburgh, April 2009. SSAISB.

[6] Joanna J. Bryson. Age-related inhibition and learning effects: Evidence from transitive performance. In *The* 31^{st} *Annual Meeting of the Cognitive Science Society (CogSci 2009)*, pages 3040–3045, Amsterdam, 2009. Lawrence Erlbaum Associates.

[7] Joanna J. Bryson. Building persons is a choice. *Erwägen Wissen Ethik*, 20(2):195–197, November 2009. commentary on Anne Foerst, *Robots and Theology*.

[8] Joanna J. Bryson. Robots should be slaves. In Yorick Wilks, editor, *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, pages 63–74. John Benjamins, Amsterdam, March 2010.

[9] Joanna J. Bryson and Jonathan C. S. Leong. Primate errors in transitive 'inference': A two-tier learning model. *Animal Cognition*, 10(1):1–15, January 2007.

[10] Timothy J. Bussey, E. Clea Warburton, John P. Aggleton, and Janice L. Muir. Fornix lesions can facilitate acquisition of the transverse patterning task: A challenge for "configural" theories of hippocampal function. *Journal of Neuroscience*, 18(4): 1622, 1998.

[11] Richard Cooper, Tim Shallice, and Jonathon Farringdon. Symbolic and continuous processes in the automatic selection of actions. In John Hallam, editor, *Hybrid Problems, Hybrid Solutions, Frontiers in Artificial Intelligence and Applications*, pages 27–37. IOS Press, Amsterdam, 1995.

[12] Richard Dawkins. *The Extended Phenotype: The Gene As the Unit of Selection*. W.H. Freeman & Company, 1982.

[13] Daniel C. Dennett. Why you can't make a computer that feels pain. In *Brainstorms*, pages 190–229. Bradford Books, Montgomery, Vermont, 1978. page numbers are from the 1986 Harvester Press Edition, Brighton, Sussex.

[14] Daniel C. Dennett. Are we explaining consciousness yet? *Cognition*, 79:221–237, 2001.

[15] Daniel C. Dennett. Can we really close the cartesian theater? Is there a homunculus in our brain? In John Dittami, editor, *Proceedings of the Vienna Conferences on Consciousness*. University of Vienna Press, 2009. in press, available from the Web.

[16] Jeffrey M. Ellenbogen, Peter T. Hu, Jessica D. Payne, Debra Titone, and Matthew P. Walker. Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences*, 104(18):7723, 2007.

[17] J. Garcia and R. A. Koelling. The relation of cue to consequence in avoidance learning. *Psychonomic Science*, 4:123–124, 1966.

[18] L. Grosenick, T.S. Clement, and R.D. Fernald. Fish can infer social rank by observation alone. *Nature*, 445:429–432, 2007.

[19] Mitch R. Harris and Brendan O. McGonigle. A model of transitive choice. *The Quarterly Journal of Experimental Psychology*, 47B(3):319–348, 1994.

[20] M. Haskell, F. Wemelsfelder, M. T. Mendl, S. Calvert, and A. B. Lawrence. The effect of substrate-enriched and substrate-impoverished housing environments on the diversity of behaviour in pigs. *Behavior*, 133:741–761, 1996.

[21] Marc D. Hauser. Perseveration, inhibition and the prefrontal cortex: A new look. *Current Opinion in Neurobiology*, 9:214–222, 1999.

[22] Katja U. Heubel, Daniel J. Rankin, and Hannah Kokko. How to go extinct by mating too much: Population consequences of male mate choice and efficiency in a sexual-asexual species complex. *Oikos*, 118(4):513–520, 2009. ISSN 0030-1299.

[23] P. N. Hineline and H. Rachlin. Escape and avoidance of shock by pigeons pecking a key. *Journal of Experimental Analysis of Behavior*, 12:533–538, 1969.

[24] Natasha Martin and Brent Alsop. Transitive inference and awareness in humans. *Behavioural Processes*, 67:157–165, 2004.

[25] Brendan O. McGonigle and Margaret Chalmers. Monkeys are rational! *The Quarterly Journal of Experimental Psychology*, 45B(3):189–228, 1992.

[26] Gerd B. Müller. Evo-devo as a discipline. *Evolving Pathways: Key Themes in Evolutionary Developmental Biology*, pages 5–30, 2008.

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

[28] Jean Piaget. *The Construction of Reality in the Child*. Basic Books, New York, 1954.

[29] Peter R. Rapp, Mary T. Kansky, and Howard Eichenbaum. Learning and memory for hierarchical relationships in the monkey: Effects of aging. *Behavioral Neuroscience*, 110(5):887–897, October 1996.

[30] Laurie R. Santos and Marc D. Hauser. A non-human primate's understanding of solidity: Dissociations between seeing and acting. *Developmental Science*, 5:F1–F7, 2002.

[31] M.N. Shadlen and W.T. Newsome. The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *Journal of Neuroscience*, 18(10): 3870–3896, 1998.

[32] M. Siemann and J. D. Delius. Implicit deductive reasoning in humans. *Naturwissenshaften*, 80:364–366, 1993.

[33] Michael Sipser. *Introduction to the Theory of Computation*. PWS, Thompson, Boston, MA, second edition, 2005.

[34] Elizabeth S. Spelke, K. Breinlinger, J. Macomber, and K. Jacobson. Origins of knowledge. *Psychological Review*, 99:605–632, 1992.

[35] Bernard Thierry. Unity in diversity: Lessons from macaque societies. *Evolutionary Anthropology*, 16:224–238, 2007.

[36] Anthony Trewavas. Green plants as intelligent organisms. *Trends in plant science*, 10(9):413–419, 2005. ISSN 1360-1385.

[37] Stuart A. West, Ashleigh S. Griffin, and Andy Gardner. Evolutionary explanations for cooperation. *Current Biology*, 17: R661–R672,, August 21 2007.

[38] M. Wheeler. Minds, things, and materiality. In Lambros Malafouris and Colin Renfrew, editors, *The cognitive life of things: Recasting the boundaries of the mind*, Cambridge, May 2010. McDonald Institute for Archaeological Research.

[39] Clive D. L. Wynne. A minimal model of transitive inference. In C. D. L. Wynne and J. E. R. Staddon, editors, *Models of Action*, pages 269–307. Lawrence Erlbaum Associates, Mahwah, NJ, 1998.

High-Dimensional Perceptual Signals and Synthetic Phenomenology

Antonio Chella[1] **and Salvatore Gaglio**[1]

Abstract. Synthetic phenomenology, in the sense of Chrisley [1], mainly focuses on the analysis of simplified perceptual signals with small or reduced dimensionality. Instead, we claim that synthetic phenomenology should be analysed in terms of dynamic perceptual signals with huge dimensionality. We claim that forms of dimensionality reduction of the perceptual signals, as done e.g. in typical robot vision applications, are characteristics of automatic "unconscious" processing. An effective "conscious" process actually deals with and must exploit the richness of the perceptual signals coming from the retina. We explore the hypothesis of a high-resolution buffer for the visual process and we discuss an application in a cognitive system for robot vision.[1]

1 INTRODUCTION

It has been questioned if robots could have qualitative, phenomenal experiences in the sense discussed by, among others, Nagel [2] and Chalmers [3]. For our present concerns, we will speak of robot phenomenology according to the synthetic phenomenology approach introduced in the seminal work of Chrisley [1].

The synthetic phenomenology approach focuses on two main efforts:

- the characterization of phenomenal states possessed or modelled by a robot;
- the use of a robot to help specifying phenomenal states.

However, the studies of synthetic phenomenology reported so far (see, e.g., Chrisley [1] and Aleksander [4]) mainly concerned the analysis of pre-processed signals with small or reduced dimensionality.

We claim that effective synthetic phenomenology should analyse raw dynamic perceptual signals with huge dimensionality generated from dynamic data directly coming from the sensory systems without using any form of compression, but instead augmenting their own dimensionality by suitable computer vision processing.

In fact, when we consider the perceptual signals coming from the retina during saccadic movements, we find a huge number of receptors that give rise to high-dimensional spatio-temporal signals over time, what Kuipers named "the firehose of experience" [5].

We claim that any forms of dimensionality reduction of the perceptual signals, as in, e.g., typical robot vision applications, are characteristics of automatic "unconscious" processing.

Effective processes relevant for synthetic phenomenology instead exploit the richness of the dynamic perceptual signals coming from the retina. To reduce the input dimensionality by compression may mean to throw the phenomenology out from the system.

2 THE COGNITIVE VISION SYSTEM

In previous papers [6-8] we presented a cognitive vision system organized in three computational *areas* – a term which is reminiscent of the cortical areas in the brain, and its relationship with experience (Figure 1).

The *subconceptual* area is concerned with the processing of data coming from the sensors. Here, information is not yet organized in terms of conceptual structures and categories. From the point of view of artificial vision, this area includes all the processes that extract suitable features, edges, and surfaces; that perform image segmentation; and that extract the 3D information of the perceived scene. In this sense, the subconceptual area is implemented as a pool of processes that tightly interact with the high-dimensional buffer (Figure 1) typically in bottom-up modalities.

In the *linguistic* area, representation and processing are based on a logic-oriented formalism. We adopt the term "linguistic" instead of the overloaded term "symbolic", because we want to stress the reference to formal languages in the knowledge representation tradition.

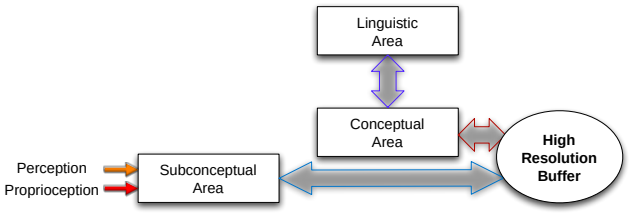

Figure 1. The cognitive vision system.

The *conceptual* area is intermediate between the subconceptual and the linguistic areas. Here, data is organized in conceptual "gestaltic" structures, that are still independent of any linguistic description. The symbolic formalism of the linguistic area is interpreted on aggregation of these structures. The conceptual area is implemented as a pool of processes that tightly interact with the high-dimensional buffer (Figure 1) typically in top-down modalities.

[1] Dipartimento di Ingegneria Informatica, Università di Palermo, Italy. Email: {chella, gaglio}@unipa.it.

In [6] we assumed that, in the case of static scenes, the conceptual area is a metric space in which each point corresponds to a 3D primitive shape, characterized according to Constructive Solid Geometry (CSG) schema. In particular, we adopted standard 3D primitives as cubes, spheres, planes, cylinders, and *superquadrics* [9] as primitives of the CSG. Therefore, a point in the conceptual area summarizes the parameters of a particular instance of a 3D primitive.

In order to represent composite objects that cannot be reduced to single 3D primitives, we assume that they correspond to groups of primitives. Figure 2 (left) shows a hammer composed of two 3D primitives, corresponding to its handle and to its head. Figure 2 (right) shows a picture of how hammers are represented in the conceptual and linguistic areas of the vision system. 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*.

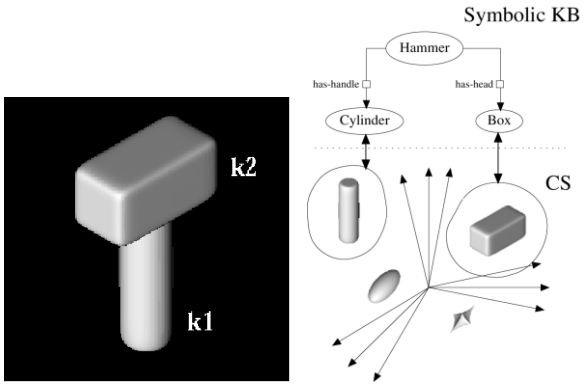

Figure 2. A hammer made of two superquadrics, and its representation in the conceptual and linguistic area.

In the described vision system, a relevant role for synthetic phenomenology is played by the high-dimensional buffer that receives information from the subconceptual area in a bottom-up modality and from the conceptual area in a top-down modality.

As described in the rest of the paper, the processes that operate over the buffer use the 3D information stored in the conceptual area and the raw data coming from sensors and processed by the subconceptual area to build a high-dimensional reconstruction of the scene, which is perceived in terms of features, boundaries, shapes, 2D and 3D information.

3 THE HIGH-DIMENSIONAL BUFFER HYPOTHESIS

Recent findings in neuroscience seem to reconsider the possible role of the V1 area in the brain. Typically (see, e.g., Marr [10]), this area has been considered as a feed-forward area able to extract local features as edges.

Evidence is discussed (Lee, Mumford et al. [11], Lee and Mumford [12]) that this area may have the functional role of a high-dimensional buffer for the lower and higher visual areas in the brain.

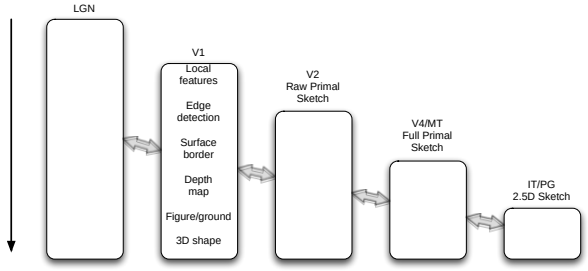

Figure 3. Interaction between the V1 area and the higher level visual areas (adapted from Lee, Mumford et al. [11]).

In this framework, the V1 area could also be involved in many higher level computations that involve "high resolution details, fine geometry and spatial precision" [11].

Figure 3 (adapted from Lee, Mumford et al. [11]) shows a possible role of the V1 area in the visual system, in which the higher level areas interact over time by means of the V1 area.

A computational model of this area should take into consideration the storage of local features and results of edge detection as in classic models of V1, but also surface borders, the depth map, figure/ground segmentation and 3D shapes of the perceived objects, necessary for the higher level visual area.

Therefore, from the computational point of view, a suitable model could be a blackboard [13] in which the different processes related to the visual areas may interact and exchange data both from bottom-up and top-down modalities.

It should be stressed that, according to this model, the data dimensionality in this area will grow according to the performed computer vision computations. Moreover, it could be the case that not all the information stored in the buffer is consistent and coherent, as in the case of the aperture problem in motion perception or the concave-convex problem in estimating shape from shading.

The information stored in the buffer $B(t)$ at time t may be viewed as a high-dimensional vector storing the information of the perceived image:

$$B(t) = \begin{bmatrix} Image \\ Edges \\ Surfaces \\ \vdots \\ 2.5D\ Shape \\ 3D\ Volume \\ 3D\ Shape \end{bmatrix}$$

The information stored in this area may be considered as a sort of "intrinsic image" of the perceived scene, i.e., a structural model of the scene [14].

This vector is high dimensional because of its several components but also because it is built up by means of fovea movements over time. In fact, the fovea acquires more and more information about the scene during saccadic movements giving rise to a high-dimensional spatio-temporal vector $B(t)$.

We hypothesize that the main focus of the synthetic phenomenology of a robot is in the high-resolution buffer accessed by bottom-up processes related to the information directly coming from the camera, and by top-down

"unconscious" processes related to the information stored in the higher level vision areas.

Therefore, no dimensionality reduction is performed, but instead, these processes augment the dimensionality of the buffer by storing the results of their computations over time.

This model resembles the Global Workspace Theory proposed by Baars (see Baars [15] for an introduction). While the computational model is similar as both models refer to a blackboard system, in our model essentially the high-resolution buffer contains the raw information coming from the camera and all the results of the computer vision processes. Therefore the buffer operates at a lower level with respect to the GWT. And in fact, the processing of the image happens spontaneously and without any form of volitions or voluntary actions.

Actually, our model is in the line of the "multiple drafts" hypothesis proposed by Dennett [16], in which different and contrasting hypotheses are generated, coexist and are destroyed over time.

The hypothesis of a high-resolution buffer inspired the promising research area of "deep machine learning" [17]. It should be noticed that, despite the obvious computational problems when dealing with high-dimensional vectors, spaces with increased dimensionality could be a munificence when searching for global minima [18]. High-dimensionality actually presents great advantages for the purposes of classification and regression, and it is at the basis of the Kernel Machines, as pointed out by Cortes and Vapnik [19].

4 A STATIC EXAMPLE

As an example of a high-dimensional buffer, we describe the experimental setup at the Robotics Lab of the University of Palermo (see [6] for details).

We have chosen a simple experimental framework that avoids some typical complex problems in computer vision. The framework consists of static scenes made up of objects like hammers, tennis balls and computer mice; all the objects rest on a uniform, visually-contrastive planar backdrop. The objects are easy to segment and they are arranged in order to avoid occlusions. Sensory data are 2D images acquired by a video camera (two-dimensional arrays of pixels) representing an orthogonal view of the observed scene, as in Figure 4.

Once again, it should be stressed that this is a simplified static example, because there are no fovea movements; we hypothesize that the whole image is immediately available to the sensors of the vision system. To exploit the whole power of the multi-dimensional buffer, we should also consider the spatio-temporal signals coming from the fovea during saccadic movements.

Starting from the sensory data of the static scene which is the 0^{th} component of $B(t)$ (see Figure 5), the region-growing process computes the segmentation map of the image. As a result, the image is initially partitioned into elementary regions of uniform brightness. This is a low-level, bottom-up process that generates the 1^{st} component of $B(t)$.

The depth map is then computed by a process that performs the shape from shading estimation, thus generating the 2^{nd} component of $B(t)$. It should be noticed that the shape from shading process receives as bottom-up input both the segmentation map and the original image and it makes the top-down assumption that the surfaces in the scene are convex ones.

Figure 4. The observed static scene.

Both the depth map and the information about the segmented regions are employed as input to the volumetric region process. The operation of this block produces the volumetric representation of the input depth map by means of a spatial array. In the current implementation, the result is a discrete representation of the spatial bulk of the objects present in the scene by voxels, i.e., in terms of primitive volume elements. This is the 3^{rd} component of the buffer $B(t)$.

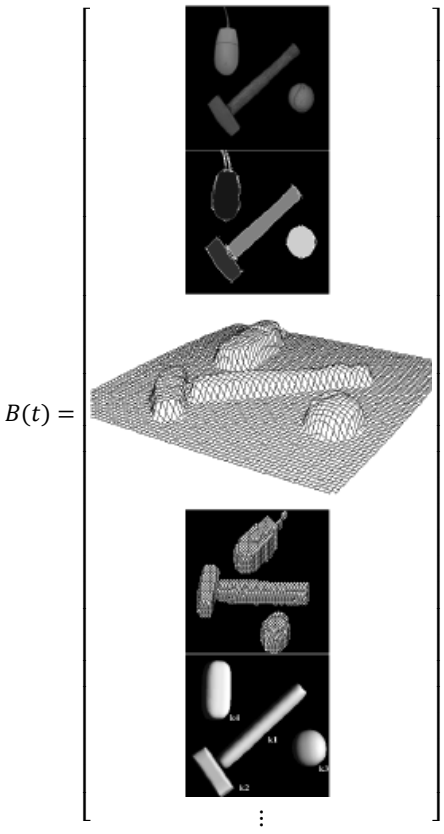

Figure 5. The multi-dimensional buffer related with the static scene in Figure 4.

The 4^{th} component of $B(t)$ in the current implementation is the fitting of the obtained volumes in terms of 3D shapes k_1, k_2, k_3, k_4. This is a bottom-up and top-down process involving the previous components of the buffer and also

suitable expectations related to the acquired scene. Its construction is an active process, driven by both the external flow of information and the inner model of the world.

As a result, the vector $B(t)$ in Figure 5 is highly multidimensional. It should be noticed that this reconstruction is viewer-dependent for some components, i.e., the 0^{th} to 3^{rd} components, and viewer-independent for the 4^{th} component. Again, not all the information stored in the buffer should be consistent and coherent.

We claim that the spatio-temporal evolution of $B(t)$ plays a fundamental role in the synthetic phenomenology of the robot, by means of foveal movements that scan the scene during time.

This buffer allows the robot to directly answer typical phenomenological questions, like, for instance, "which is the part of that object and which are its colour and brightness at that point in space? How fast is it moving? Is it alerting you?", and so on.

5 CONCLUSIONS & FUTURE WORK

This paper claims that the synthetic phenomenology of a robot should take into account the whole huge richness of signals coming from the sensors exploited by the vision processes occurring in the higher vision area.

The main hypothesis is that a high-dimensional vector that acts as a buffer for low and high level vision processes is the main site for exploiting the synthetic phenomenology of the robot.

We illustrated the idea of a multi-dimensional buffer with reference to a simplified case of a static scene without fovea movement.

Future works will concern the full exploitation of the spatio-temporal multi-dimensional buffer considering fovea movements, as in the work of Chrisley [1].

A related research line will concern the analysis of the huge dimensionality of the resulting vector by means of suitable kernel machines [20]. The goal of this approach is to get advantages of the high dimensionality of the resulting buffer vector in order to classify and to generate anticipations of the perceived scene by means of linear operators of low complexity, but without dealing with the high complexity related with the dimensionality of the resulting vector.

REFERENCES

[1] R. Chrisley, Synthetic Phenomenology, *International Journal of Machine Consciousness*, **1**, 1, pp. 53 – 70 (2009).

[2] T. Nagel, What is it like to be a bat? *Philosophical Review*, 83, pp. 435–450 (1974).

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

[4] I. Aleksander, *The World in My Mind, My Mind in the World: Key Mechanisms of Consciousness in Humans Animals and Machines*, Imprint Academic, Exeter, UK (2005).

[5] B. Kuipers, Drinking from the Firehose of Experience, *Artificial Intelligence in Medicine*, **44**, pp. 135 – 170, 2008.

[6] A. Chella, M. Frixione and S. Gaglio, A cognitive architecture for artificial vision, *Artificial Intelligence*, 89, pp. 73–111 (1997).

[7] A. Chella, M. Frixione and S. Gaglio, Understanding dynamic scenes, *Artificial Intelligence*, 123, pp. 89-132 (2000).

[8] A. Chella and S. Gaglio, In Search of Computational Correlates of Artificial Qualia, in: B. Goertzel, P. Hitzler, M. Hutter (eds.): *Artificial General Intelligence, Proc. of the Second Conference on Artificial Intelligence AGI-09*, Atlantis Press, Amsterdam, pp. 13 – 18 (2009).

[9] A. Jaklič, A. Leonardis and F. Solina, *Segmentation and Recovery of Superquadrics*. Kluwer Academic Publishers. Boston, MA (2000).

[10] D. Marr, D. *Vision*. W.H. Freeman, New York (1982).

[11] T.S. Lee, D. Mumford, R. Romero and V. Lamme, The Role of the Primary Visual Cortex in Higher Level Vision, *Vision Research*, **38**, 2429 – 2454 (1998).

[12] T.S. Lee and D. Mumford, Hierarchical Bayesian Inference in the Visual Cortex, *J. Opt. Soc. Am. A*, **20**, 7, pp. 1434 – 1448 (2003).

[13] N. Carver and V. Lesser, Blackboard systems for knowledge-based signal understanding. In *Symbolic and Knowledge-Based Signal Processing*. Oppenheim, A. V., and Nawab, S. H. (eds.). Prentice-Hall, Englewood Cliffs, N. J. pp. 205 – 250 (1992).

[14] J.M. Tenenbaum, M.A. Fischler and H.G. Barrow, Scene Modeling: A Structural Basis for Image Description, *Computer Graphics and Image Processing*, **12**, pp. 407 – 425 (1980).

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

[16] D. Dennett, D., *Consciousness Explained*. Penguin, London (1993).

[17] I. Arel, D. C. Rose, T. P. Karnowski, Deep Machine Learning - A New Frontier in Artificial Intelligence Research, *IEEE Computational Intelligence Magazine*, pp. 13 – 18, November 2010.

[18] R. Hecht-Nielsen, The Munificence of High-Dimensionality, in: I. Aleksander, J. Taylor (eds.), *Artificial Neural Networks, 2*. Elsevier, Amsterdam, pp. 1017 – 1031 (1992).

[19] C. Cortes, V. Vapnik, Support Vector Networks, *Machine Learning*, **20**, pp. 273 – 297 (1995).

[20] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge (2000).

Information Integration, Data Integration and Machine Consciousness

David Gamez[1]

Abstract. Information integration is a property of systems of connected elements that expresses the extent to which they are capable of entering a large number of states that result from causal interactions among their elements. In recent years a number of people have claimed that there is a link between information integration and consciousness and a number of algorithms for measuring information integration have been put forward. This paper gives an overview of the conceptual and experimental issues surrounding information integration and explores some of the links between information integration and machine consciousness.

1 INTRODUCTION

Neuroscience research often uses information measures, such as mutual information or transfer entropy, to identify the anatomical, functional and effective connections between different parts of the brain or a neural simulation [1, 2]. Many authors have observed that integration is a key feature of conscious states [3, 4], and it has been a natural progression to suggest that measures of functional and effective connectivity could be used to identify the parts of the brain that are highly integrated and thus correlated with conscious states. Tononi [5, 6] has gone beyond this correlation approach to claim that consciousness actually *is* integrated information, and proposed algorithms for identifying the areas of maximum information integration in a system [7, 8].

One of the key attractions of information integration theories of consciousness is that they are precise enough to be experimentally tested. For example, Tononi's most recent theory [5] can predict the areas of a system that are associated with consciousness, the amount of consciousness that is present and the qualitative character of this consciousness for each state of a system. This precision of information integration theories points the way towards a more scientific approach to consciousness, in which falsifiable predictions made by different mathematically formulated theories of consciousness can be systematically compared (see Section 4). A second advantage of information integration theories is that they can be applied to both artificial and natural systems: if a link could be established between information integration and consciousness in humans, then it would be possible to make convincing predictions about the consciousness of artificial systems as well.

While information integration theories are promising, there are many issues that need to be addressed. Some conceptual difficulties surrounding the nature of information and integration are covered in Section 3, and there are a number of practical problems with measuring information integration, including the performance of current algorithms, how accuracy can be evaluated, and the selection of a particular level of the system for analysis. These practical issues are examined in Section 4, which gives an overview of how information integration theories could be experimentally tested. Some of the potential applications of information integration algorithms are covered in Section 5.

2 BACKGROUND

Research on information integration and consciousness is closely linked to work on the identification of functional and effective relationships between neurons and neuron groups using neural complexity [9, 10], transfer entropy [11] and other measures. There has been some research comparing neural complexity measures and graph theory [12], and these measures have been used by a number of people to examine the anatomical, functional and effective connectivity of biological networks, either using scanning or electrode data, or large-scale models of the brain. One example of this type of work is Honey et al. [13], who used transfer entropy to study the relationship between anatomical and functional connections on a large-scale model of the macaque cortex, and demonstrated that the functional and anatomical connectivity of their model coincided on long time scales. Another example is Brovelli et al. [14], who used Granger causality to identify the functional relationships between recordings made from different sites in two monkeys as they pressed a hand lever during the wait discrimination task. Information-based analyses have also been used to guide and study the evolution of artificial neural networks connected to simulated robots [15, 16]. An overview of this type of research can be found in [1, 2].

A number of people have suggested that there is a link between information integration and consciousness [3, 4], or that information integration actually *is* consciousness [5, 6], and several algorithms for calculating information integration have been put forward. These include neural complexity [10], stateless Φ [8], state-based Φ [7],[3] causal density [17], liveliness [18], and an information integration measure that can be applied to time series data [19]. Seth et al. [20] gives a review of earlier work, identifies a number of weaknesses in Tononi and Sporns' [8] method and criticizes the link between information integration and consciousness.

There has been a limited amount of experimental work on the link between information integration and consciousness. For example, Lee et al. [21], made multi-channel EEG recordings from eight sites in conscious and unconscious subjects and constructed a covariance matrix of the recordings on each frequency band that was used to identify the complexes within

[1] Dept. of Computing, Imperial College, London SW7 2BT, UK. Email: dgamez@imperial.ac.uk.

[3] The information integration measure put forward by Tononi and Sporns [8] will be referred to as "stateless Φ" to distinguish it from the related state-based measure of Φ of Balduzzi and Tononi [7].

the 8 node network using Tononi and Sporns' [8] algorithm. This experiment found that the information integration capacity of the network in the gamma band was significantly higher when subjects were conscious. Massimini et al. [22, 23] have carried out experiments in which a TMS pulse was applied to the subject's brain and the resulting activity was recorded using EEG. Massimini et al. found that the activity resulting from the TMS pulse was more localized and less differentiated when the subjects were unconscious, suggesting that a combination of integration and differentiation is linked to conscious states. In the simulation work on information integration detailed predictions have been made about the amount and distribution of consciousness in an 18,000 neuron network [24], and some preliminary comparisons have been carried out between state-based Φ and liveliness [18].

3 INFORMATION OR DATA INTEGRATION?

3.1 The Nature of Information

> Information is notorious for coming in many forms and having many meanings. It can be associated with several explanations, depending on the perspective adopted and the requirements and desiderata one has in mind.
>
> Floridi [25], p.1

In the work using information theory to measure brain connectivity, the exact nature of information is not important because Shannon's information equations are used as mathematical tools to identify the functional and effective connections between groups of neurons. However, the nature of information does become important when a property of the system called information integration is linked to consciousness. In this shift it is no longer the *relationships* between biological neurons that are important, but the *presence* of information integration in the system. It then becomes necessary to say what information is and how it can be identified in an arbitrary system.

To understand the problem of identifying information, imagine that an aubergine is lying on the table in front of you. Simply through its existence on the table, this aubergine contains the information that there is an aubergine on the table in front of you. Cut the aubergine open and a pattern of seeds are revealed, which can be interpreted as letters in a particular language. The genetic code of the aubergine can also be read and the aubergine contains an enormous amount of information about the location of each of its atoms relative to a particular reference point. Each of these types of information at different levels of the aubergine can be transformed into other types of information. For example, the sequence of nucleotides in the aubergine's DNA could be remapped into a sequence of numbers representing an image or sound.

This example of the aubergine suggests that physical objects can be interpreted as containing a virtually infinite quantity of information, which is relative to an observer and level of abstraction that define the information states in a system. A theory of information that can handle this apparent relativity is the General Definition of Information (GDI). According to Floridi's [26] formulation of the GDI, σ is an instance of information, understood as semantic content, if and only if:

GDI.1) σ consists of n data, for *n* >= 1;
GDI.2) the data are well-formed;
GDI.3) the well formed data are meaningful.

This GDI is based on the notion of *data*, which Floridi describes as a lack of uniformity in the world. The data that is accessible to us and can be read depends on pre-theoretical differences in the physical world, called *dedomena*, which cannot be experienced without an interpretation that is applied to the world. These *dedomena* are the conditions of possibility for experienced data - something like Kant's *noumena* or Locke's *substance* - that make the differences that we can measure and manipulate possible. For example, *dedomena* might make the measureable difference between higher and lower charge in a battery possible, and this type of measureable difference between physical states can in turn be used to create higher levels of data, such as symbols. The interface that defines the scope and type of data in a system is called a level of abstraction by Floridi.

From the point of view of the information integration theory of consciousness, this distinction between *dedomena* and data that we understand and manipulate is important because only *dedomena* can be considered to be an objective property of the system. The data or sets of differences that we actually extract will always be the result of a particular interpretation. This is not a problem if we are interested in correlations between information integration and consciousness, but it is an issue for Tononi's claim that information integration *is* consciousness because consciousness is typically thought to be an objective feature of the world, not a subjective interpretation of a system by an observer. One potential way around this problem would be to say that consciousness is the integration between differences in a physical attribute of the system, such as its electric field, which is thought to be more than just a subjective interpretation.

Once a method for identifying data in a system has been defined, the next stage is to specify a syntax that will enable *well-formed* data to be extracted in a systematic way. For example, the sequences of nucleotides that constitute the aubergine's genetic code can only be read when we can distinguish between sequences coding proteins and junk DNA. If we want to interpret the aubergine's seeds as Arabic writing, the seed pattern will have to conform to the shapes of the letters in Arabic, the order of the letters will have to conform to the orthography of Arabic and the order of the words will have to conform to the grammar of Arabic.

While it is relatively easy to extract well-formed data from a system, this data might may not be *meaningful* in any way. Suppose that the aubergine is in half, covered with a grid of millimetre squares and a 1 is read off if the square contains an even number of seeds and a 0 is read off if the square contains an odd number of seeds. This sequence of 1s and 0s is well formed data because it conforms to a specified syntax, but since it lacks meaning, it is not information according to the GDI. The question of what makes data meaningful is much more difficult than the identification of data in a system, and Floridi's approach is to describe semantic content as a combination of data and queries. So, for example, the proposition "The earth only has one moon" can be interpreted as a piece of meaningful data in which the semantic content is the question "Does the Earth only have one moon?" and the answer "yes" is a single bit of data.

Starting with the work of Shannon [27], there has been an extensive amount of work on the communication of information, describing the entropy of an information source, the mutual information between two devices and the maximum rate of

communication over a channel. However, as Floridi points out, Shannon's mathematical theory of communication (MTC) is a theory about data transmission, not about information transmission, because it does not take the meaning of the messages into account: "since MTC is a theory of information without meaning (not in the sense of meaningless, but in the sense of not yet meaningful), and since we have seen that [information – meaning = data], 'mathematical theory of data communication' is a far more appropriate description of this branch of probability theory than 'information theory'."([26], p. 33). This suggests that "information integration" algorithms based on Shannon's work, such as [8], are actually measures of *data* integration, unless it can be shown that the integrated data carries semantic content. This distinction between data and information poses a separate question about whether there is a link between information integration and consciousness, where information is understood as meaningful data. This is particularly relevant when considering embodied theories of consciousness, since meaningful data could be data that co-varies with the world.

3.2 Integration of Information

There are many ways of interpreting the notion of integrated information, including data fusion, meta data about information and statistical and causal relationships between items of information. The type of information integration that is claimed to be linked to consciousness has a very specific meaning because it is not just the *integration* that is important, but the *differentiation* of the information states as well. Tononi [5, 6] illustrates this idea of differentiated integration using the example of a digital camera sensor with a million photodiodes. This sensor is highly differentiated because it can enter $2^{1,000,000}$ different states but in each of these states the photodiodes are acting independently and there is no integration between them. In contrast, consider a million Christmas lights connected to a single switch: when the switch is on, the lights are on; when the switch is off, the lights are off. In this system there is a high level of integration between the switch and the lights, but almost no differentiation because the system can only enter two possible states: all lights on or all lights off. In between the camera photodiode and the Christmas lights are systems that are both differentiated and integrated: they can enter a large number of different states and these states are the result of causal interactions between the elements. According to Tononi [5, 6], a key example of differentiated and integrated systems are the areas associated with consciousness in the human brain.

The algorithms that have been put forward for measuring information integration - for example [7, 8] - are intended to quantify the balance between differentiation and integration in a system of connected elements. While algorithms for measuring the integration between items of information are relatively straightforward – for example, statistical or causal measures – and the differentiation of a system can be quantified using information entropy, it is a challenging task to find an algorithm that can quantify the combination of differentiation and integration. Some of the performance and accuracy issues raised by the current algorithms are covered in Section 4.3.

4 TESTING INFORMATION INTEGRATION

4.1 Introduction

Information integration is an empirical theory about a link between a measured feature of the physical world and phenomenal experience. Its great strength is that it makes strong claims about the world that can be shown to be false. Many other theories of consciousness, such as higher order thought [28], might be thought to be intuitively plausible, but they are not scientific if they cannot be experimentally tested.

Experiments on the link between information integration and consciousness are likely to involve the following steps:

1. Select a system that is known to be conscious or commonly agreed to be conscious.
2. Measure the information integration of the system.
3. Measure the consciousness of the system.
4. Identify correlations between information integration and consciousness.
5. Test predictions made by information integration about the consciousness of the system.

The following sections cover each of these stages in more detail. Although this discussion is framed in terms of the information integration theory of consciousness, a similar approach could be applied to any theory of consciousness that is expressed in a precise mathematical or algorithmic form.

4.2 The Platinum Standard System

To establish whether information integration in a physical system is linked to conscious states it is necessary to start with a physical system that is known or commonly agreed to be associated with consciousness. Although we typically assume that infants and higher mammals are conscious, the only system that is confidently associated with consciousness is the awake normal adult human brain. By 'normal' it is meant that the brain is undamaged and its functions and measurements fall within two standard deviations for the human species. 'Awake' is intended in a non-technical sense to indicate that the brain is functioning in a way that is typically considered 'conscious'. This type of wakefulness is distinct from the medical definition, since apparently wakeful states can be exhibited by people in a vegetative state who are unlikely to be conscious [29]. While there will be times when the awake normal adult human brain is not conscious – for example, epileptic automatism [30] - a science of consciousness has to start somewhere, and the awake normal adult human brain is the physical system that we are most certain is typically associated with conscious states.

The awake normal adult human brain will be referred to as the platinum standard system. Just as a platinum-iridium standard bar in Paris was used to define the length of a metre, the awake normal adult human brain is our platinum standard for a conscious system. If this physical system is not associated with conscious states most of the time, then nothing is. A further assumption, that the consciousness associated with different platinum standard systems is roughly the same, may become necessary for detailed predictions about the contents of consciousness.

It is important to note that artificial systems cannot be used to test the link between information integration and consciousness

because it is not clear whether they are associated with conscious states. If a link between information integration and consciousness could be established, then it would become possible to make predictions about the consciousness of artificial systems using information integration, but this link has to be demonstrated on a platinum standard system first.

4.3 Measuring Information Integration in the Physical Platinum Standard System

To investigate the link between information integration and consciousness it is necessary to measure the amount of information integration in the platinum standard system. The first stage is the definition of the level of abstraction that will be used in the experiments. Data in the brain can be defined at many different levels – for example, sub-atomic, atomic, molecular, neural or neuron group - and it is far from clear whether different levels of abstraction will lead to different amounts of information integration in the system, or whether the levels will coincide. As an example, consider the problem of measuring colour in a sack of oranges. If colour is measured at the level of individual oranges, then the sack of oranges will be pronounced orange. Likewise, an analysis at the level of segments will result in an orange colour, but analyses at the sub-atomic level or at the level of pips will result in zero orange colour. Within information integration analyses, the key experimental challenge is to identify whether the levels coincide or contradict – for example, whether information integration analyses at the level of ions match analyses based on areas of the brain. There is also a challenging question about whether the interpretation of the neural code will affect the amount of information integration – for example, rate-based analyses could give very different results from analyses based on polychromous groups [31].

The presence of multiple information integration algorithms that apparently measure the same objective property of the physical world raises the question about which is the most *accurate* algorithm and how this accuracy can be measured. The accuracy of information integration algorithms could be evaluated by making the (problematic) assumption that information integration is correlated with consciousness, and carrying out experiments – for example, using fMRI or EEG – that measure the correlation between the output of the information integration algorithms and the reports of conscious states from the platinum standard system. If information integration is correlated with consciousness, then the algorithms that most accurately predict consciousness would be the most accurate measures of information integration. The main problem with this approach is that some or all of the algorithms might be measuring a property of the brain that is correlated with consciousness, but which has nothing to do with information integration. There is also the issue that our spatial and temporal access to the brain is severely limited, which makes it very difficult to measure information integration in humans.

Another way of measuring the accuracy of information integration algorithms is to create simulated networks with regions that we expect to have high information integration – for example, a neural network with several highly intra-connected modules would be expected to have higher information integration within the modules. Different information integration algorithms could be run on these networks and their output compared with the areas of expected maximum information integration. This approach has the problem that our intuitions about the areas of maximum information integration might not be correct, but there does not appear to be a way of measuring the information integration of a network that does not depend on a particular algorithm. While the simulated networks approach is problematic, until our access to the brain improves it appears to be the only method available for the comparison of different measures of information integration.

A second problem with measuring information integration is the performance of some of the current algorithms - for example, it has been predicted to take 10^{9000} years to fully analyze an 18,000 neuron network using Tononi and Sporns' algorithm [32] and it could take 100 million years to analyze a network of 30 elements using Balduzzi and Tononi's algorithm [18]. Some work on addressing this issue has been carried out by Aleksander and Gamez [18], who developed an algorithm for measuring information integration based on liveliness that scales linearly with the number of neurons and connections. However, even with these improvements, current supercomputers are likely to struggle with a full analysis of the platinum standard system, which has around 10^{10} neurons and 10^{15} synapses. Although artificial systems cannot be used as platinum standard systems, they are an ideal test environment for benchmarking the performance and accuracy of different ways of measuring information integration.

A final problem with some of the current algorithms, such as state-based Φ and liveliness, is that they rely on knowledge about the underlying causal structure of the system. This is not a problem with artificial systems, where the causal structure is usually known, but typical measurements of the platinum standard system, such as fMRI, EEG or electrode data, are a sequence of states whose causal relationship is unknown. This is not an issue for Seth's Granger causality measure [33], and it could be partially addressed for the other measures by inferring the causal structure from the sequence of states using Granger causality, transfer entropy or mutual information.

4.4 Measuring Consciousness in the Platinum Standard System

In experiments on the link between information integration and consciousness, the information integration of the physical system is compared with measurements of conscious states. In the platinum standard system, consciousness is typically measured through first person reports, although other behaviours can be used to infer the presence of consciousness and its contents. It is also possible to use cognitive abilities that are systematically linked to consciousness to reliably infer the presence of consciousness in a platinum standard system [34]. There are numerous problems with the measurement of consciousness, such as the dependence on potentially fallible memory and the limited bandwidth and accuracy of human language – see [34] for a more detailed discussion.

4.5 Correlations between Information Integration and Consciousness in the Platinum Standard System

While falsifiable predictions are the gold standard for scientific theories (see Section 4.6), initial work on the link between information integration and consciousness is likely to focus on the identification of correlations between information integration and consciousness. Early experiments are likely to study whether the presence of consciousness is correlated with higher information integration; eventually research will move on to examine whether different degrees of consciousness are linked to different amounts of information integration and whether the contents of consciousness vary in the way suggested by information integration theories.

4.6 Predictions about Consciousness in the Platinum Standard System

> ...the real test of a scientific theory of consciousness is its ability to make falsifiable predictions: I shall certainly admit a system as empirical or scientific only if it is capable of being tested by experience. These considerations suggest that not the verifiability but the falsifiability of a system is to be taken as a criterion of demarcation ... I shall require that its logical form shall be such that it can be singled out, by means of empirical tests, in a negative sense: it must be possible for an empirical system to be refuted by experience.
>
> Popper [35], p.18

A substantial amount of the current work on consciousness is based on theories that are felt to be more or less intuitively plausible. To become truly scientific, the study of consciousness has to move towards a situation in which predictions made about the consciousness of a platinum standard system are compared with the platinum standard system's behavioural reports about its consciousness.

The information integration theory that is most capable of making predictions [5] can predict the areas of a physical system that are associated with consciousness, the amount of consciousness in these areas and the qualitative character of this consciousness. When our ability to measure the human brain has increased its spatial and temporal resolution (and if the performance of current algorithms can be improved), it should become possible to make predictions about the consciousness of a platinum standard system and compare these predictions with first person reports. The information integration theory of consciousness would become widely accepted if it could make large numbers of accurate predictions about the contents of consciousness of human subjects using only physical information about the system.

4.7 The Path Ahead

The problems identified in this paper currently make it impractical to systematically test information integration algorithms on platinum standard systems. Instead, initial work in this area is likely to use artificial neural networks, possibly embodied in robots, to develop more efficient algorithms and investigate novel ways of analyzing networks for information integration. This work on artificial systems will feed into experimental work using scanning data with low spatial and/or temporal resolution, such as the experiments discussed in Section 2. During this research, the predictions that are made about the consciousness of artificial systems will *not* be believed because a link will not have been established between information integration and consciousness on the platinum standard system. Eventually it is hoped that improvements in the speed of computers, increased spatial and temporal accuracy of brain scanning, and greater efficiency of information integration algorithms will make it possible to establish whether there are systematic correlations between information integration and consciousness in the platinum standard system.

If empirical evidence could establish that information integration is systematically linked with consciousness in the platinum standard system, then information integration could be used to make predictions about the consciousness of other systems. For example, we could use information integration to make believable predictions about the consciousness of artificial systems, infants and animals. We could also look back retrospectively at the systems that were analyzed for information integration in the past, and believe that these systems were conscious to the degree predicted because the information integration theory will have been rigorously proved on the platinum standard system.

It is possible that empirical research will demonstrate that information integration is *not* correlated with consciousness in the platinum standard system, or that it fails to make accurate predictions about consciousness. In this case, information integration theories should be abandoned and better approaches sought.

5 APPLICATIONS

Accurate predictions about consciousness in humans have many applications, such as measuring the degree of consciousness in coma patients, identifying whether a person is unconscious during an operation, and it might be possible for paraplegic patients to use the predicted contents of their consciousness to control artificial limbs. Accurate predictions about consciousness would also raise serious privacy issues - for example military and police interrogators could read a suspect's mind, and people's intentions are often used to identify their degree of criminal guilt: the difference between murder and manslaughter is largely a matter of intention. Predictions about animal consciousness could be used to minimize animal suffering – it might even become possible to genetically engineer food animals with little or no consciousness. Predictions about the consciousness of early embryos would have applications in abortion legislation.

Within work on machine consciousness, an accurate algorithmic measure of consciousness could be used to measure the degree to which an artificially conscious system has been constructed. It could also determine whether artificially conscious systems are suffering – one of the objections to machine consciousness raised by Metzinger [4] is that it amounts to the creation of a race of retarded infants for experimentation. This worry about the suffering and confusion of artificial systems is becoming more pressing because scanning technologies are developing to the point at which it may soon become possible to get exact data about the location and connections of every neuron in a mouse brain [36, 37]. This

would enable the real-time simulation of a particular mouse's brain, which might be capable of experiencing the same pain as the original mouse.

There has been a substantial amount of discussion of the possibility that people could scan their brains into a computer and achieve a form of digital immortality [38, 39]. After a person dies their brain would be preserved and a succession of very thin slices would be scanned and integrated together to build up a complete picture of their neurons and connections.[4] This information would then used to build a neural network with the same neurons and connections, and in theory this could be accurate enough to produce the same global behaviour. Since the person's brain will have developed in a close relationship with their body, it might be necessary to connect the simulated brain to their original body, perhaps using electrodes attached to nerves in the spinal column.

People who pay for this procedure might only be interested in perpetuating the *external behaviour* of their brain after their death, which might be possible if the neural simulation works in real time, is accurate enough, and is connected to a suitable body. However, a key question will remain as to whether this simulation of a person's brain will be as conscious as the person was before their death, or whether the simulation will be a zombie that just replicates the person's external behaviour. If information integration or a similar theory could be shown to make accurate predictions about consciousness, then it could be used to predict the extent to which a simulation of a person's brain is as conscious as the brain was before the person's death. It is possible that some simulations of a person's brain will not be conscious at all - perhaps because they are running on a time-sliced computer - so predictions about the consciousness of digitized brains might make people think very carefully about the way in which they want to be 'uploaded' after their death.

6 CONCLUSIONS

This paper has given an overview of some of the conceptual and experimental issues surrounding work on the possible link between information integration and consciousness. Information integration theories of consciousness are interesting because they open up the possibility of making predictions about consciousness, which could open up a new chapter in the scientific study of natural and artificial minds.

While artificial systems cannot be used to test the link between information integration and consciousness, they can play a key role in improving and understanding the current algorithms and addressing questions about their performance and accuracy. If a link between information integration and consciousness could be proved, then it would become possible to use information integration to make believable predictions about the consciousness of animals, artificial systems and simulated scanned brains. Although information integration is a promising approach, we are only just beginning to explore mathematical and algorithmic theories of consciousness, and experimental work may show that information integration is poorly correlated with consciousness. In this nascent field many different approaches may have to be tried before we discover a high performance high accuracy scientific theory of consciousness.

[4] It might eventually become possible to make a detailed scan of a person's brain before their death, but this is well beyond the reach of current technology.

ACKNOWLEDGEMENTS

This work was supported by a grant from the EPSRC (EP/F033516/1).

REFERENCES

[1] O. Sporns, D. R. Chialvo, M. Kaiser, and C. C. Hilgetag. Organization, Development and Function of Complex Brain Networks. *Trends Cogn Sci,* 8: 418-25 (2004).

[2] O. Sporns. (2011). *Scholarpedia Article on Brain Connectivity*. Available: http://www.scholarpedia.org/article/Brain_connectivity

[3] B. J. Baars, *A Cognitive Theory of Consciousness*. Cambridge University Press, Cambridge England ; New York (1988).

[4] T. Metzinger, *Being No One : The Self-Model Theory of Subjectivity*. MIT Press, Cambridge, Mass. (2003).

[5] G. Tononi. Consciousness as Integrated Information: A Provisional Manifesto. *Biol Bull,* 215: 216-42 (2008).

[6] G. Tononi. An Information Integration Theory of Consciousness. *BMC Neurosci,* 5: 42 (2004).

[7] D. Balduzzi and G. Tononi. Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. *PLoS Comput Biol,* 4: e1000091 (2008).

[8] G. Tononi and O. Sporns. Measuring Information Integration. *BMC Neurosci,* 4: 31 (2003).

[9] G. Tononi, G. M. Edelman, and O. Sporns. Complexity and Coherency: Integrating Information in the Brain. *Trends Cogn Sci,* 2: 474-84 (1998).

[10] G. Tononi, O. Sporns, and G. M. Edelman. A Measure for Brain Complexity: Relating Functional Segregation and Integration in the Nervous System. *Proc Natl Acad Sci U S A,* 91: 5033-7 (1994).

[11] T. Schreiber. Measuring Information Transfer. *Physical Review Letters,* 85: 461-464 (2000).

[12] M. Shanahan. Dynamical Complexity in Small-World Networks of Spiking Neurons. *Physical Review E,* 78, (2008).

[13] C. J. Honey, R. Kotter, M. Breakspear, and O. Sporns. Network Structure of Cerebral Cortex Shapes Functional Connectivity on Multiple Time Scales. *Proc Natl Acad Sci U S A,* 104: 10240-5 (2007).

[14] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler. Beta Oscillations in a Large-Scale Sensorimotor Cortical Network: Directional Influences Revealed by Granger Causality. *Proc Natl Acad Sci U S A,* 101: 9849-54 (2004).

[15] O. Sporns and M. Lungarella. Evolving Coordinated Behavior by Maximizing Information Structure. In: *Artificial Life X: Proceedings of the 10th International Conference on the Simulation and Synthesis of Living Systems*, pp. 322-329 (2006).

[16] A. K. Seth and G. M. Edelman. Environment and Behavior Influence the Complexity of Evolved Neural Networks. *Adaptive Behavior,* 12: 5-20 (2004).

[17] A. K. Seth. Cognitive Networks in Simulated Neural Systems. *Cognitive Neurodynamics* 2: 49-64 (2008).

[18] I. Aleksander and D. Gamez. Informational Theories of Consciousness: A Review and Extension. In: *Proceedings of BICS 2010*, Madrid (2010).

[19] A. B. Barrett and A. K. Seth. Practical Measures of Integrated Information for Time-Series Data. *PLoS Comput Biol,* 7: e1001052 (2011).

[20] A. K. Seth, E. Izhikevich, G. N. Reeke, and G. M. Edelman. Theories and Measures of Consciousness: An Extended Framework. *Proc Natl Acad Sci U S A,* 103: 10799-804 (2006).

[21] U. Lee, G. A. Mashour, S. Kim, G. J. Noh, and B. M. Choi. Propofol Induction Reduces the Capacity for Neural Information

Integration: Implications for the Mechanism of Consciousness and General Anesthesia. *Conscious Cogn,* 18: 56-64 (2009).
[22] F. Ferrarelli, M. Massimini, S. Sarasso, A. Casali, B. A. Riedner, G. Angelini, G. Tononi, and R. A. Pearce. Breakdown in Cortical Effective Connectivity During Midazolam-Induced Loss of Consciousness. *Proc Natl Acad Sci U S A,* 107: 2681-6 (2010).
[23] M. Massimini, M. Boly, A. Casali, M. Rosanova, and G. Tononi. A Perturbational Approach for Evaluating the Brain's Capacity for Consciousness. *Prog Brain Res,* 177: 201-14 (2009).
[24] D. Gamez. Information Integration Based Predictions About the Conscious States of a Spiking Neural Network. *Consciousness and Cognition,* 19: 294-310 (2010).
[25] L. Floridi, *Information : A Very Short Introduction*. Oxford University Press, Oxford (2010).
[26] L. Floridi. Philosophical Conceptions of Information. *Lecture Notes in Computer Science*, 5363: 13-53 (2009).
[27] C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal,* 27: 379–423, 623–656 (1948).
[28] D. M. Rosenthal. Two Concepts of Consciousness. *Philosophical Studies* 49: 329-59 (1986).
[29] S. Laureys, S. Antoine, M. Boly, S. Elincx, M. E. Faymonville, J. Berre, B. Sadzot, M. Ferring, X. De Tiege, P. van Bogaert, I. Hansen, P. Damas, N. Mavroudakis, B. Lambermont, G. Del Fiore, J. Aerts, C. Degueldre, C. Phillips, G. Franck, J. L. Vincent, M. Lamy, A. Luxen, G. Moonen, S. Goldman, and P. Maquet. Brain Function in the Vegetative State. *Acta Neurol Belg,* 102: 177-85 (2002).
[30] V. S. Ramachandran and S. Blakeslee, *Phantoms in the Brain : Probing the Mysteries of the Human Mind*, 1st ed. William Morrow, New York (1998).
[31] E. M. Izhikevich. Polychronization: Computation with Spikes. *Neural Comput,* 18: 245-82 (2006).
[32] D. Gamez. *The Development and Analysis of Conscious Machines*. PhD, Department of Computer Science, University of Essex, 2008.
[33] A. K. Seth. Causal Networks in Simulated Neural Systems. *Cogn Neurodyn,* 2: 49-64 (2008).
[34] M. Shanahan, *Embodiment and the Inner Life : Cognition and Consciousness in the Space of Possible Minds*. Oxford University Press, Oxford (2010).
[35] K. R. Popper, *The Logic of Scientific Discovery*. Routledge, London (2002).
[36] M. Helmstaedter, K. L. Briggman, and W. Denk. 3d Structural Imaging of the Brain with Photons and Electrons. *Curr Opin Neurobiol,* 18: 633-41 (2008).
[37] A. Li, H. Gong, B. Zhang, Q. Wang, C. Yan, J. Wu, Q. Liu, S. Zeng, and Q. Luo. Micro-Optical Sectioning Tomography to Obtain a High-Resolution Atlas of the Mouse Brain. *Science,* 330: 1404-8 (2010).
[38] R. Kurzweil, *The Age of Spiritual Machines : How We Will Live, Work and Think in the New Age of Intelligent Machines*. Phoenix, London (1999).
[39] D. Chalmers. The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies,* 17: 7-65 (2010).

A model of primitive consciousness in autonomously adaptive system under a framework of reinforcement learning

Yasuo Kinouchi[1], Yoshihiro Kato[1], Hiroki Hayashi[1], Yusuke Katsumata[1], Kazuhisa Kitakaze[1], and Shoji Inabayashi[2]

Abstract. A model of primitive consciousness is proposed based on the investigation of a system, composed of stochastic neural networks, that autonomously adapts without a teacher to its environment. This system must not only respond to the environment as fast as possible but also provide a response that is appropriate to the situation based on its previous experiences. The system should grasp the situation, decide the appropriate action, and adapt by modifying its own configuration on the basis of its experience. To do these things quickly and efficiently without a teacher as a single entity, the system adapts to its environment based on a reinforcement learning framework in the wide sense, and has one evaluation mechanism based on rewards or punishments in the system itself.

The system is composed of six modules: a perception module, an integration module that calculates a candidate for an action, a motor control module, an episodic memory module, a working memory module, and a basic control module which has an evaluation mechanism influenced by emotion.

One of the main functions of the system is temporal memorization of signals that represent 'states' and 'values' (predicted rewards) to determine the successive system action. The states are composed of information that describes the situation in the environment and the system itself at that time. A candidate for an action is calculated in the integration module under conditions given by temporarily maintained information. This function corresponds to an action decision based on policy in reinforcement learning.

If we assume there is a filtering mechanism such that we can only perceive the temporarily maintained information in the system, our phenomenal consciousness corresponds to a logical space composed of the main signals in reinforcement learning. Moreover, 'the evaluation mechanism and its states' functions as a self in the space. As a whole, it is shown that primitive consciousness is a compact logical space necessary for autonomous adaptation of the system.

1 INTRODUCTION

A model of primitive consciousness is proposed on the basis of results of an investigation of a system which is composed of stochastic neural networks that autonomously adapts without a teacher to its environment. This system must not only respond to the environment as fast as possible but also provide a response that is appropriate to the situation based on its previous experiences. The system should grasp the situation, decide the appropriate action, and adapt by modifying its own configuration on the basis of its experience. To do these things quickly and efficiently without a teacher as a single entity, the system adapts to its environment based on a reinforcement learning framework in the wide sense, and has one evaluation mechanism based on rewards or punishments in the system itself.

In the modelling of consciousness, it is important to note the difference between phenomenal consciousness and functional consciousness.[1] To clarify the difference, a model with two layers, a physical layer and a logical layer, is adopted.[2],[3] All signals are processed in detail by the neural nodes in the physical layer. By contrast, the minimum information necessary for the system to adapt itself, selected from the physical layer, composes the logical layer or space. The operations in the logical layer are represented by interactions between only the selected information.

The physical layer of the system is composed of six modules. The perception module recognizes concepts from micro-features of sensor outputs. The integration module quickly calculates a candidate for an action as a whole system. In this module each interconnected node corresponds to one equation which represents constraints in the action selection. Whole nodes in the module constitute a set of simultaneous equations that are solved by an iterative method. The other four modules are for motor control, episodic memory, basic control, which has an evaluation mechanism affected by emotion, and for working memory.

One of the main functions of the system is temporal memorization of signals that represent 'states' and 'values' (predicted rewards) to determine the successive system action. The states are composed of information that describes the situation in the environment and the system itself at that time. A candidate for an action is calculated in the integration module under conditions given by temporarily maintained information. This function corresponds to an action decision based on policy in reinforcement learning. The maintained information is transferred to the episodic memory module, memorized for a long period, and recollected when necessary.

In the conventional model of reinforcement learning, the main signals: 'states, policy, and value' are formed by something that is not the system itself. [4] However, in the autonomously adaptive system they should be formed inside the system by the system itself in the process of adaptation. If we assume there is a filtering mechanism such that 'we can only perceive information in the temporal memory', our phenomenal consciousness corresponds to a logical space composed of the main signals in reinforcement learning. Moreover the evaluation mechanism and its states functions as a self in the space. As a whole, it is shown that primitive consciousness is a compact logical space that the system needs for autonomous adaptation.

1 Dept. of Information Systems, Tokyo Univ. of Information Sciences, Japan. kinouchi@rsch.tuis.ac.jp
2 Pacific Technos Corp, Japan

2 BASIC CONDITIONS OF AUTONOMOUSLY ADAPTIVE SYSTEM

We assumed the basic conditions of an autonomously adaptive system as shown below.
(i) The system must autonomously adapt to a complex environment without a teacher. Large amounts of information are input into the system from this environment.
(ii) To adapt autonomously to the environment, the system has self-action-decision functions to adapt to certain circumstances and a learning-control that varies the system configuration itself based on an inherent value-evaluation mechanism, such as a reward and punishment.
(iii) The system has basic inherent automatic or semi-automatic functions that correspond to the instinctual body-control functions of animals.
(iv) Artificial neural nodes with stochastic characteristics, in which information is represented by random pulse frequency of activated nodes, are implemented.
(v) The system must decide as quickly as possible upon an appropriate or suitable action in accordance with it's current situation.
(vi) The system must learn from its own experiences that consist of the sequential perceptions, actions, and rewards in the environment as quickly and effectively as possible.
(vii) The system must reduce as much of its resources, such as the nodes, and energy used in the system itself as possible.

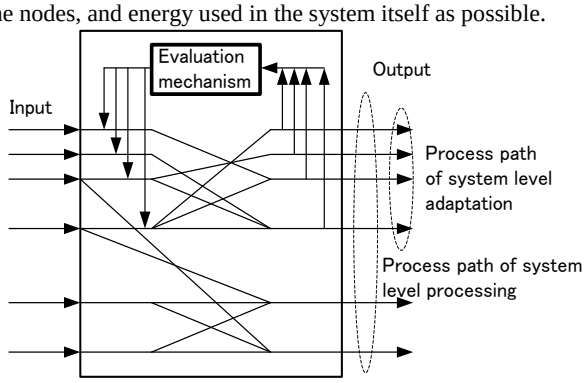

Figure1. Schematic diagram of process path of system level processing and adaptation

3 MODELING METHOD

3.1 System level and non-system level processing

To ensure the system operates quickly and appropriately, the processing of the system is classified into two flows, system level and non-system level. In system level processing, the action decision consists of the whole system; for example the decision of approach or avoidance of the system, are done. To make these decisions, various pieces of perceived information need to be gathered from a wide area within the system and interact. The path lengths of the signal flows lengthen. Additionally, if the system processes this information on a first-come-first-serve basis, this means that important system information is not given a higher priority. Various types of information need to be buffered in a certain period from which important information needs to be selected. On the other hand, in non-system level processing, various motor signals are directly controlled by the corresponding local or restricted information. System level processing needs more time and resources than non-system level processing. We assume that system level processing is activated only for situations where the system has to act as a whole system.

3.2 Levels of adaptation

Adaptation of the system is executed by two mechanisms shown below.
(i) A system level adaptation mechanism: Adaptation as a whole system based on evaluation of rewards and punishment
(ii) A node level adaptation mechanism: Adaptation through variation of neural nodes, such as increasing or decreasing of the weights of nodes.

These two mechanisms are not exclusive. In many cases these mechanisms operate concurrently. System level adaptation can be executed by only a part of the system level processing mechanism. Moreover, information in this adaptation (i) has to be interrelated to each other. As shown in Figure 1, to select an appropriate action in successive operation after that time, all paths in system level adaptation have to run commonly under the influence of the unique result of evaluation.

3.3 Phenomenal consciousness and functional consciousness

We feel the following items in daily life.
(a) Real-time information of the real space, situation, and our body as the actual phenomena experienced outside our brains.
(b) Recollected or thinking imageries.
(c) States of mind such as a mental or emotional phenomenon inside our brain

To concretely grasp these feelings in an autonomously adaptive system, a screen of images representing the real world, screen of images of recollected contents, and states of emotion are modeled in physical configuration as operations of the neural nodes. However, when modeling consciousness, the difference between phenomenal and functional consciousness should be noted. To clarify the difference, a model with two layers, a physical one and a logical one, was used.

The physical layer is represented by the physical operations of neural nodes. In contrast, the logical layer is represented as a logical space that consists of information selected and mapped from the physical layer through the virtualizing method of information technology.[4],[5]

3.4 Adaptation under a framework of reinforcement learning

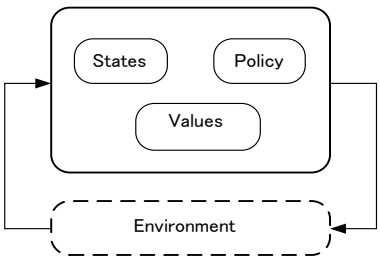

Figure 2.
Framework of reinforcement learning

A schematic diagram of reinforcement learning composed of states, policy, and values is shown in Figure 2, and under a

framework of reinforcement learning, the whole system adaptation is modeled as shown in Figure 3. Here, states consist of information that originated from real stimulation, information based on mental images, information arising from the inner system, such as system conditions, and information from the evaluation, resulting from previously evaluation at the discrete time, t.

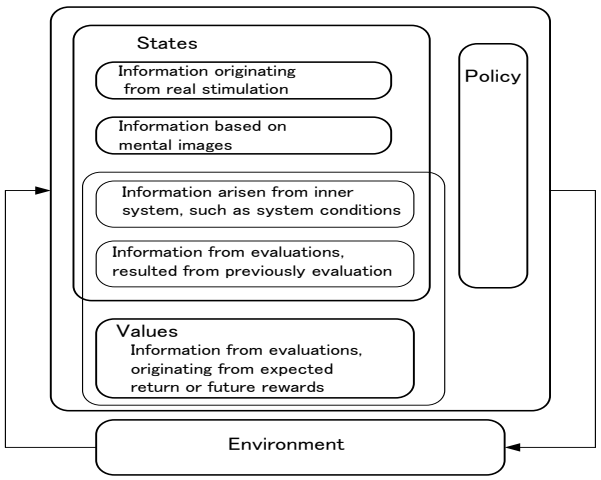

Figure 3. Information of autonomously adaptive system under a framework in the reinforcement learning

A policy is used to calculate the candidate of an action performed at that time and based on these states. Values consist of information from evaluations, originated from expected return or future rewards. The system must maximize the values. Rewards or punishment gained at time t varied the states of the system at time t+1.

Although in the usual engineering model of reinforcement learning, the main signals: states, policy, and values are formed by something that is not the system itself, in an autonomously adaptive system these should be formed inside the system by the system itself in the process of adaptation.

By using neural networks, the above model is configured in detail as a kind of sequential circuit with inner states composed of states and values, and a function which calculates a candidate action as policy based on the inner states. Inner states are maintained by temporal buffering, or memorized in short-term or long-term memory.

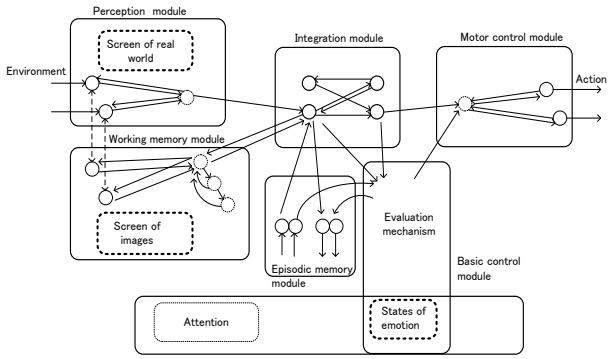

Figure 4. Configuration of physical layer

4. PHYSICAL CONFIGURATION

As shown in Figure 4, the system consists of six modules.

(1) The perception module consists of micro-feature nodes, concept perception nodes and the screen of the real world as shown in Figure 5. The micro-feature nodes represent corresponding sensor signal mainly from environment.[7] The concept perception nodes recognize concepts based on related micro-features. Output signals of the module, representing activated concepts at that time are maintained by mutual stimulation between corresponding concept perception nodes and micro-features on the screen for a short time. The screen of the real world is depicted by micro-features that are active at that time. (In this paper, what information is necessary to depict a screen is investigated, but how the screen is depicted is not referred to.)

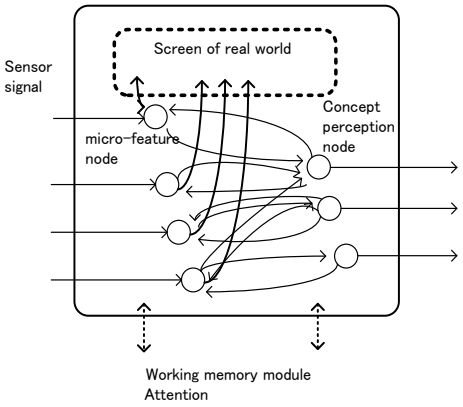

Figure 5. Configuration of the perception module

(2) The integration module is composed of massive calculation nodes corresponding to concepts or actions. These nodes are interconnected to each other as shown in Figure 6.[8] The module quickly calculates a desired state that includes not only candidate selection for an action but also situation recognition as a whole system based on the information input from a screen of images, the screen of the the real world, and states of emotion. A candidate for an action is transferred to the motor control module. This function corresponds to an action decision based on 'policy' in reinforcement learning. The recognized situation represented as information of some important object at that time is transferred to the working memory module and episodic memory module. Each interconnected node, in many cases corresponding to a concept, represents constraints as one equation, and whole nodes in the module constitute a set of simultaneous equations that are solved by using an iterative method. Constant terms of the simultaneous equations are given from other modules as inputs.

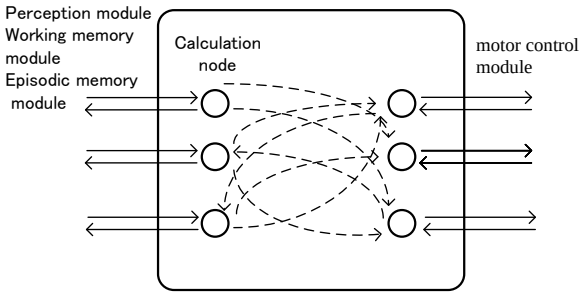

Figure 6 . Configuration of the integration module

(3) The motor control module transmits orders to the motor for action.
(4)The episodic memory module sequentially memorizes a group of information that corresponds to a screen of the real world, a screen of images, and emotional states including the results of the evaluation.
(5) The working memory module has image micro-feature nodes, image concept nodes and the screen of images as shown in Figure 7. Each image micro-feature node corresponds to the micro-feature node with the same attribute in the perception module. Each image concept node represents the same concept corresponding to the concept node in the perception module, and has functions of delayed memory that memorize multiple states at time t, t-1,...,t-n. First, output signals from the integration module stimulate the image concept node. Then the node stimulates image micro-feature nodes that belong to the image concept node. When these image micro-feature and concept nodes have sufficient support signals from the basic control module, nodes activation are maintained by mutual stimulation between nodes for a short time, similar to the perception module. Images on the screen are depicted by these activated image micro-feature nodes, and the states of image concept nodes are transmitted to the integration module, as input at time t+1. Moreover, stimulating delayed memory from basic control module, the states of the image screen and image concept nodes are changed to previous states at time t-1,...,t-n. The working memory module controls calculations in the integration module by varying the states of image concept nodes connected to calculation nodes in the integration module.

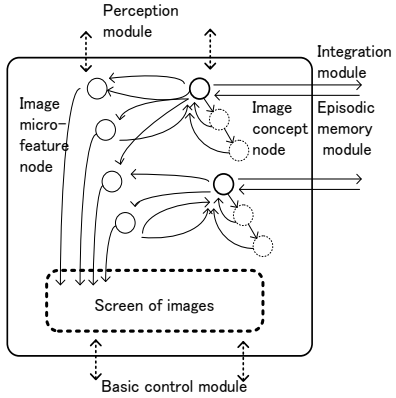

Figure 7 . Configuration of working memory module

(6) The basic control module has three important functions, such as evaluating action with the integrated module, processing path control by opening and shutting gates, and maintaining emotional states.

5. MODELING OF CONSCIOUSNESS AND SELF

5.1 Model of consciousness in a logical space

As shown in Figure 8, a group of information composed of screen of the real world, screen of images, and states of emotion forms the main inner states of the system, and dominantly decides the system's successive action in the physical configuration. This means that the system can be controlled by the group of information which corresponds to our daily feeling as phenomenal conscious experiences.

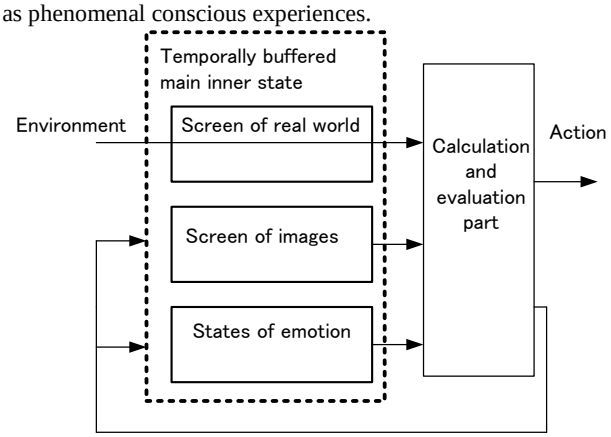

Figure 8. System operation as a sequential circuit

To model phenomenal consciousness clearly, a logical space based on the virtualization method that is implemented in usual information systems is adopted. This logical space is composed of only the group of information: screen of the real world, screen of imageries, and states of emotion. Interrelationships between the information in the group in the logical space are defined based on the physical connection of neural nodes. However, in the logical space, the operation of each node is invisible. Functions are executed as black boxes and are not affected by the physical locations of neural nodes.

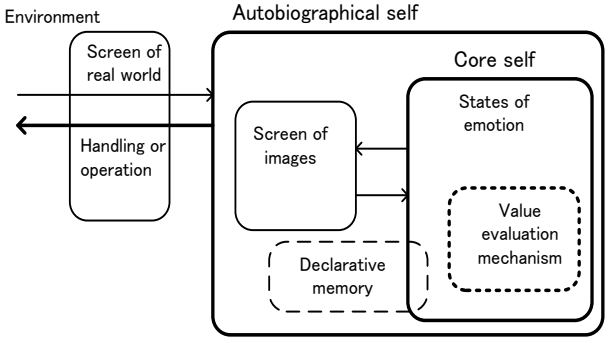

Figure 9. Schematic diagram of autobiographical and core self in logical layer

5.2 Model of the self

To model the self in the logical space, areas A and B are introduced. At each time, only one area, either A or B, is effective. These areas include the evaluation mechanism of a whole system, and the evaluation mechanism evaluates and represents interests of elements inside of each area. The elements outside the area are not evaluated, and have information on what it is and where are its location and constituent parts.

Each area decides an action or handles an element outside the area to maximize the interests inside it. This operation introduces the subjectivity of each area. Each area needs information on location to handle elements outside the area. However, as the area does not handle the elements inside it, information on locations that fall within it are not needed. As a result, the self is not related to physical location, it is at the center of a logical space. We assume that the area A corresponds to the core self, and area B corresponds to the autobiographical self as shown in Figure 9. [9]

6. CONCLUSIONS

We propose a model of primitive consciousness and self on the basis of an investigation of a system that adapts to its environment based on a reinforcement learning framework.

Primitive consciousness in the autonomously adaptive system functions only when the system works based on the system level adaptation in system level processing. To efficiently process and adapt, the following are used: a group of information: the screen of the real world, the screen of images, and the states of emotion including the results of evaluation. The information maintained temporally in the perception, working memory, and basic control modules dominantly decides the next action. Additionally, the information is memorized in the long-term declarative memory, recollected to the working memory when needs arise, and is used to take the appropriate action. In the system level adaptation, all paths have to interrelate through evaluation results that is unique information in the system.

Phenomenal consciousness is clarified in the logical space that consists of a group of information. This logical space is defined by selection and mapping from the operation of the physical nodes on the basis of the virtualizing method in the information system. We assume that our daily experiences, to feel or not to feel, execute this virtualizing, and when we feel, the results of evaluation are always included.

The self is explained as a function of the area in which the evaluation mechanism is included. The evaluation mechanism gives priority to interests in the area. The area is used to maximize the interests in the area, and handles elements outside of the area. These operations generate a subject-object relationship. As a whole, it is shown that 'primitive consciousness' is a compact logical space needed to allow the system to autonomously adapt to environment.

We are now preparing to simulate the system on a computer to clarify its operational characteristics in more detail.

ACKNOWLEDGMENTS

We thank Prof. Pentti O. Haikonen of the Philosophy Department, University of Illinois at Springfield, Prof. Igor Aleksander of Imperial College, and Dr. Kenji Itoh of the Graduate School of Medicine, University of Tokyo for their useful suggestions.

REFERENCES

[1] S. Franklin, S. D'Mello, B. Baars, and U. Ramamurthy, Evolutionary Pressures for Perceptual Stability and Self as Guides to Machine Consciousness, International Journal of Machine Consciousness 1(1), (2009)

[2] Y. Kinouchi, A logical model of consciousness on a neural network system with a simple abstract brain-like structure in Proc. NOKIA Workshop on Machine Consciousness (Helsinki, 2008) pp.22-26

[3] Y. Kinouchi, A logical model of consciousness on an autonomously adaptive system, IJMC Vol. 1, No. 2 (2009)

[4] R. Sutton, and A. Barto, Reinforcement Learning An Introduction, The MIT Press, London, England,(1998)

[5] J. Gray, and Reuter, A. Transaction processing: concepts and techniques. Morgan Kaufmann (1993)

[6] D. Norman, Invisible Computer, MIT Press (1998).

[7] P. Haikonen, Robot Brains-Circuit and Systems for Conscious machines, Wiley, West Sussex (2007).

[8] Y. Kinouchi ,S.Inabayashi, Y.Nakazaki, A model of primitive consciousness on an autonomously adaptive system, ASSC14 (2010)

[9] A. Damasio, Descartes' Error: Emotion, Reason, and the Human Brain, Putnam's Sons, (1994).

Development and situated cognition: Preconditions for machine consciousness

Riccardo Manzotti[1]

Abstract. First, I discuss whether our current understanding of the related notions of embodiment, embeddedness, and situatedness is satisfactory. Second, I consider a stronger set of requirements that may flesh out a stronger model of situated cognition and, tentatively, of situated consciousness. Finally, I discuss a robotic implementation and try to show the kind of intimate agent-environment relationship that legitimises cognitive extension.

1 WHAT IS SITUATED COGNITION? A DEVELOPMENTAL PERSPECTIVE

Although there has been a widespread upsurge of interest as to the notion of embodied, situated and extended cognition [1-7], it is still unclear which conditions must obtain for cognition, and possibly consciousness, to spread into the environment. In this regard, the available cognitive models do not always offer completely consistent and overlapping criteria [6, 8-12].

Any robotic system is apparently embedded and situated in the environment. A robot has a body and it does things in response to actual events. However, it may be argued that such a loose characterization obtains also for a washing machine or for a thermostat. Are they embodied and situated in the same way as much more complex agents like ants, monkeys, and human beings? It is fair to answer negatively. The relevant kind of situatedness and embodiment must consist in something else.

However, if a body and sensory-motor capacities are not enough, what else then? Situatedness entails an intimate coupling between the agent cognitive functions, the body structure and the environment. But what is exactly such an intriguing *intimate* coupling? When does it occur? Is the extent to which an agent is cognitively situated quantifiable? Here, I venture to consider a working definition based on a model for situated consciousness – a kind of radical phenomenal externalism whose ontological skeleton has been outlined elsewhere [13-15] envisaging a kind of *vehicle externalism* [16-17]. In this paper, the issue of what constitutes a significant coupling between the world and a cognitive agent is addressed in order to check whether it is relevant to consciousness or not. The gist of the intuition at hand is to consider embodiment and development together.

Put simply, I suggest that *an agent is situated in a given environment to the extent that its cognitive structures are the result of developing inside that environment*. The word *situated* is used in opposition to *embodied* and *embedded* because the agent is cognitively and phenomenally spread beyond those events that fall within the physical boundary of the body.

[1] Institute of Behaviour and Communication "G.P.Fabris", IULM University, via Carlo Bo, 8, 20143, Milan, Italy.
Email: riccardo.manzotti@iulm.it.

In this paper, a criterion to single out the conditions that characterize the unity between the environment and the agent is suggested. Development and individual history may be the missing ingredients. Unfortunately, development is a very loose notion. Here, development refers to more pervasive changes than those usually encompassed by learning (for instance changing one's long term goals rather than learning a better way to achieve them). Consider a robot whose behavioural patterns are fixed at design time. Notwithstanding its efficiency, will such a robot be situated? I'd answer negatively. A truly situated agent keeps changing its cognitive structure. The issue at stake is whether there is a particular way to adapt that endorses a situated agent. This is not a philosophical discussion as to the true nature of 'development'. Rather it is important to focus on the actual conditions that foster the proper kind of coupling between environment and agent.

The appeal to development might seem reasonable but unnecessary. It may be argued that while the coupling between an agent and its environment is likely to be higher if the agent has developed within that environment, the link is by no means necessary. For example, the Furby toy (http://en.wikipedia.org/wiki/Furby) is programmed to speak less "Furbish" and more English over time, but this "development" happens entirely independently of its environment. A robot or animal with fixed behaviour patterns could be highly coupled with certain environments, and highly decoupled from others.

Yet the outcome of the Furby development is totally independent from the environment in which the Furby operates. If the Furby were raised in a Chinese speaking environment, it would not speak more Chinese – the Furby will speak more English. Irrespective of its actual environment, the Furby develops in a preprogrammed way. So, while it is true that in the proper environment the Furby would match the surrounding environment, it is nevertheless true that its coupling is not the result of its individual history. On the contrary, in biological beings, it is well known that most of their development is open, in the sense that it is not constrained by innate rules. It is customary to distinguish between two kinds of development: ontogenesis and epigenesis. The former is similar to that of the Furby and is largely independent of the actual environment, while the latter is 'open' and heavily caused by the actual interactions during development.

However, the above considerations suggest the following picture. On one hand, development may or may not be necessary to produce environmental coupling. On the other hand, coupling may not be enough to get situated cognition. That is why it could be worthwhile to consider the possibility that situatedness requires both environmental coupling and development. Each of them may be insufficient.

If the behavioural structure is caused by the environment, it may be argued that the individual history is somehow *constitutive* of the resulting agent. If the individual history of an agent has no causal efficacy, it cannot be considered as constitutive of the agent. A certain kind of 'open' development may thus be taken

into consideration as the process during which the bonds with the environment are established. Furthermore, there are plenty of causal accounts of information, meaning, phenomenal experience, content, symbol grounding, and the like [15, 18-20]. They require the occurrence of actual causal links between the environment and the agent.
Unfortunately, it appears rather difficult to provide feasible and quantifiable methods to measure either coupling or development. As a result, scholars often prefer to focus on the current informational state internal to the system under scrutiny rather than to deal with the causal history that led to that state of the agent. Furthermore, the state of an agent is often considered separately from the structure of the agent, as if the internal state and its control structure were two independent aspects. This separation is questionable as the human brain shows. In biological beings such a separation is clearly a mistake. There aren't memory banks or CPUs in the brain.
These considerations ought to shed some light on the issue of the feasibility of a measure of coupling and situatedness.
There may be conceptual and empirical arguments suggesting to avoid considering only the *internal* data mesh of a cognitive architecture. Similarly it could be misleading to consider only a measure of agent-environmental coupling. In fact, the developmental story might be something more than a way of producing high levels of coupling. Coupling itself might not be sufficient. If the goal is to understand and indeed replicate a mind (cognitive and phenomenal), both the proper kind of development and environmental coupling might be necessary in order to constitute a situated agent.

2 A MODEL FOR EMBODIMENT BASED BOTH ON CAUSAL ENTANGLEMENT AND ON DEVELOPMENT

In this section, an architecture aiming at situated cognition and consciousness is outlined. The architecture does not pretend to be either conclusive or experimentally satisfying. However, it is a cognitive architecture that has been partially implemented in previous setups [21-22] and partially presented in previous works [23-24]. The architecture aims at implementing the kind of development and environmental coupling briefly mentioned in the previous section. Or, at least, it ought to suggest how such issues are to be addressed. Hopefully the model makes predictions on what phenomenal content could be.
At the root of the architecture there is a causal structure that can be replicated again and again at different levels of complexity. The architecture will span three levels: the unit level, the module level, and the architecture level. In fact, what we describe here is not the blueprint of an architecture, but rather a recipe to generate a cognitive architecture by means of interacting with the environment. Another aspect, which should raise some hope of success, is that there is no separation between control, data, and structure – everything contributes to the whole and there is no way to separate control from data.
Consider a body with multiple actuators and sensors, each sensor providing multiple incoming channels. At the onset, the overall structure is unknown. Is there any suggestion how to generate automatically a cognitive architecture capable of somehow mastering the sensori-motor contingencies of the forthcoming agent? Instead of presenting the outline of an architecture, it would be useful to have a sort of meta-architecture for generating an autonomously environment-coupled agent. This is a very ambitious goal that it is not going to be solved here. However, this paper aims to go in such a direction.
Informally, the proposed recipe to generate the actual architecture is the following. Suppose one has a hive of elementary units capable of becoming dedicated to specific events in the environment. Suppose also that such hives are able to play the role both of pattern recognition units and of controller of further actions and learning. In other words, suppose that the traditional separation between data, control, and goals is set aside. Finally, suppose that such units may be combined freely in order to single out higher level external stimuli, to map higher level motor patterns, and to pursue higher level goals. This would be an ideal recipe to generate a situated architecture, since the resulting architecture would be totally and causally dependent on the environment and the individual history.
In this section, I try to outline a sketch of a tentative implementation inspired by the above-mentioned ideal recipe by means of three steps: the elementary unit, the intentional module, and the intentional architecture.
The Elementary Unit. The first element to define is the unit capable of picking up a stimulus from the surrounding environment and then becoming dedicated to it. This unit will be the basic element of the developing architecture. It is a unit receiving an input (whether it be as simple as a bit or as complex as any data structure you could envisage) and producing an output (a scalar, an integer or a logical value) -- from many to one, so to speak. As we will see, the unit also has a control input, but this is a detail that will be specified below.
The main goal of the unit is getting matched with an arbitrary stimulus and, after such a matching, having the task of being causally related with this stimulus. It could be implemented in many different ways. Formally, it can be expressed by an undefined function waiting for its first input before being matched to it forever. Basically, it is like having a selective gate that is going to be burned on its first input. After being "burned", the unit has a significant output only if the current input resembles the first input. If a continuous similarity function and a continuous output are used, the gate tunes the passage of information rather than blocking/allowing it, yet the general principle remains the same.
The final detail is the control input signal. Due to the irreversible nature of the matching, it could make sense to have a way to signal when the first input is received. Since the unit could be activated only at a certain moment, it is useful preserving its potential until certain conditions are obtained. Thus there is a need for an external control signal that will activate the unit signalling that the first input is on its way.
Due to its very simple role, this unit may be dubbed the *elementary* unit. It is important that, up to now, the unit is not committed to any particular kind of data or internal implementation. The unit is potentially very general. It can be adapted to any kind of inputs: characters, words, numbers, vectors, images.
A more formal description will help getting the gist of the unit. Any kind of input domain C can be defined. The output domain may be conveniently defined to be a real number in the 0 to 1 interval. Finally, a similarity function has to be chosen – at worst, the identity function could be used. The similarity function fs:C×C→ [0,1] takes two arguments from the input domain and its output must be maximum if and only if the two arguments are equal. The more they are different, according to any feasible criteria, the more the output of such function must get closer to the minimum. The similarity function is used to implement the elementary unit internal function F_u:C→[0,1], which is

the function that will change its behaviour forever after its first input. To recap, the elementary unit behaviour is open in the sense that is undefined until it gets coupled with a specific external stimuli. In this way, such a unity will become forever coupled with an event in the individual history and development of the architecture. F_u is defined as follows:

$$F_u(s_t) = \begin{cases} 0 & t < t_0 \\ 1 & t = t_0 \\ fs(s_t, s_{t_0}) & t > t_0 \end{cases}$$

It must be stressed that t_0 is an arbitrary chosen instant marked by the external control signal mentioned above.

F_u is a function waiting for something to happen before adopting its final and fixed way of working. It entails that the unity does not do anything until it gets coupled with some unpredictable and incoming external stimulus s_{t0}. Once it happens, that unity will forever provide the maximum output for whatever future stimulus identical with that seminal incoming stimulus that shaped its behaviour forever. The idea is to have a unity that is going to get causally coupled with some aspect of the environment. The output is provided by similarity function that will forever have as one of its arguments the stimulus s_{t0}. Such a function may be as simple as the identity function or as complex as the designer likes.

Consider a few examples. Suppose that the incoming domain C is constituted by alphabetic characters, that the similarity function fs:C×C→[0,1] is the identity function, and that the output domain is the binary set {0,1}. The function F_u is thus complete and the elementary unit can be implemented. The function F has no predictable behaviour until it receives the first input. After that it will output 1 only when a character identical to the one received at its beginning is received. Imagine that a possible input is the following sequence of characters: 'S', 'T', 'R', 'E', 'S', 'S'. The output will then be 1, 0, 0, 0, 1, 1.

A simple variation on the similarity function would permit a slightly fuzzier notion of similarity: two characters are similar (output equal to 1) either if they are the same or if they are next to each other in the alphabetical order. Given such a similarity function and the same input, the output would now be 1, 1, 1, 0, 1, 1. To make things more complex, a continuous output domain such as [0,1] is admitted and the similarity function is changed to $fs(c)=1-AD/TC$ where AD is the alphabetical distance and TC is the total number of alphabetical characters. With the above input, the output would then be: 1, 0.96, 0.96, 0.65, 1, 1. Clearly, everything depends on the first input.

A useful formalization of the unit is the one having vectors as its input domain. Given two vectors $\boldsymbol{v}$,$\boldsymbol{w}$, a simple way to implement the similarity function is using a normalized version of a distance function between vectors $d(\boldsymbol{v},\boldsymbol{w})$. Suitable candidates are the Minkowski function or the Tanimoto distance [25]. Using greyscale images as input vectors, in the past one of the author used the following correlation function [22]:

$$fs(\vec{v},\vec{w}) = \frac{1}{2}\left[1 - C(\vec{v},\vec{w})\right] = \frac{1}{2}\left[1 - \frac{\sum (v_i - \mu_v)(w_i - \mu_w)}{\sqrt{\sum (v_i - \mu_v)^2 \cdot \sum (w_i - \mu_w)^2}}\right]$$

These are just a few of the examples that could be made. It is important to stress that the unit is as simple as it is very adaptable and open to many different domains and implementation. Furthermore, the unit has a few features that are worth being stressed.

First, the unit does not distinguish between data and processing. In some sense, it is a unit of memory since its internal function is fixed on a certain input thereby keeping a trace of it. The stimulus s_{to} is somehow stored in the unit. However, it is not an explicit memory, since there is not a stored value, but rather a variation in its behaviour by means of its internal function. Another interesting aspect is that the unit shows a behaviour which is the result of the coupling with the environment. When the unit is "burned", it is also forever causally and historically matched to a certain aspect of the environment.

There is no way to predict the future of the unit's behaviour since it is the result of the contingent interaction with the environment. If the input is unknown, the unit behaviour is unknown too. The unit behaviour is not hardwired in any sense.

Finally, the unit seems to mirror, to a certain extent, some aspect of its own environment without having to replicate it. Slightly more philosophically, the unit enables the existence of a pattern in the environment by allowing it to produce effects through itself (more detailed considerations on this issue are outlined in [14].

By itself the elementary unit could seem pretty useless. However things get more interesting once a large number of them are assembled together.

The intentional module. Suppose one has the capability of implementing and packing many elementary units into the same physical or logical package, putting together a hive of units. The result could be a structure here labelled as the *intentional module*. This module has already been put to the test in a very simplified robotic setup aiming at developing new motivations and controlling the gaze of a camera towards unexpected classes of visual stimuli [22, 26].

The simplest way to step from the elementary unit to the module is to pack together a huge number of elementary units all receiving the same input source. To avoid them behaving exactly the same, some mechanisms that prevents them from being shaped by the same input at the same time must be added. There are various ways to do this. A simple way is to number them and then to require the units to be burned sequentially. This could be obtained by means of the external control signal each elementary unit has. Because of its importance the external control signal is labelled the *relevant signal*. The name expresses that such a signal is relevant in the life of each elementary unit since it controls to which input value the unit is matched forever.

Since the burning of each elementary unit will surely have a cost in terms of resources, it could make sense to add some more stringent conditions before switching on a relevant signal (and thus coupling forever an elementary unit to a certain input). Which conditions? To a certain extent they could be hard-wired and thus derived from some a priori knowledge the designer wants to inject into the system. Or, if the system is the result of some earlier generations of similar systems, it could be derived from the past history of previous generations. But it is definitely interesting whether the system could develop its own criteria to assign resources to further incoming inputs. By and large, the size of the input domain is surely much larger than that that can be mapped by the available elementary units. In short, a feasible solution is having two explicitly divided sets of criteria working concurrently and then adding their outputs together so as to have a relevant signal to be sent to the elementary units. The first set could be a set of hardwired functions trying to pin down external conditions that, for one reason or another, could be relevant for

the system. The second set should be somehow derived from the growing set of matched elementary units themselves.
On the basis of what information should these two sets of criteria operate? A simple solution is on the basis of the incoming information with an important difference. The hardwired criteria could operate straight on the incoming data since they are hardwired and thus apply off-the-shelf rules. On the other hand, the derived set of criteria could use the output of the elementary units thereby using a historically selected subset of the incoming signals.

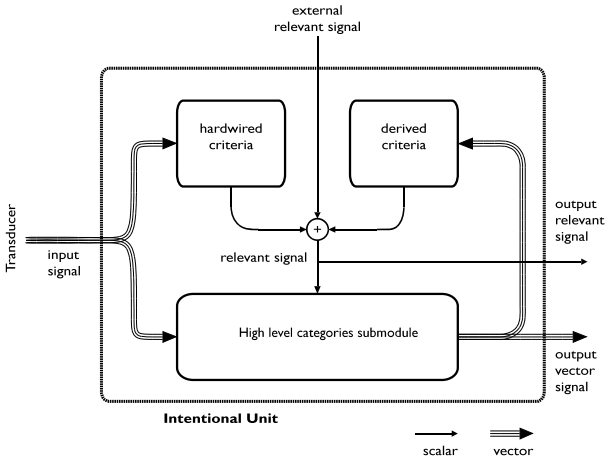

Figure 1. An intentional module.

Shifting from the logical structure to the level of implementation, the above-mentioned elements are packed into three submodules as shown in Figure 1. First the elementary units are packed into a huge array. Second, the hard-wired criteria are grouped together in such a way that they receive the signals jointly with the array of elementary units. Third, there is a third sub-module grouping together the criteria derived by the past activity of the elementary units. Moreover, the module receives an external relevant signal that flanks the two internally generated ones. The reason for this will get clearer below. The output relevant signal is extremely important since it compresses all the past history of the system with respect to the current input signal. The vector output is the result both of the history and of the hard-wired criteria. Each element of this vector is the output of an elementary unit. Therefore the output vector has as many elements as there are burned and activated elementary units. The value of each element expresses how much the corresponding elementary unit is activated by the current input signal, which in turn means how much the current input signal is similar to a given past input.
Formally the intentional module implements a function

$$F_m\left(r,\vec{v}\right)=\begin{pmatrix} r_n \\ \vec{v}_n \end{pmatrix}$$

where r is the external control signal, $\vec{v}$ is the input vector, r_n is the output relevant signal, and $\vec{v}_n$ is the output vector signal. Given an array of N elementary units, the output vector is:

$$\vec{v}_n=\begin{pmatrix} F_u^1\left(r,\vec{v}\right) \\ \dots \\ F_u^N\left(r,\vec{v}\right) \end{pmatrix}$$

The intentional architecture. The above intentional module is sufficient to implement classic conditioning, attentive behaviour, and a rough self-generation of new goals [14]. Furthermore, the module can process any kind of data. It does not have to know in advance whether the incoming data is originated by images, sounds, texts, or whatever. The module embeds its history in its structure. The module is unpredictable since its behaviour is the result of a tight coupling with the environment.
An aspect that is worth of some consideration is that the module receives a vector and a scalar and it outputs a vector and a scalar as well. As a result it is possible to use the intentional module as the building block of a much larger architecture.
Now we will outline how to exploit these three features in order to design a more complex architecture with a robotic implementation in mind. Consider the case of a robot moving in an environment such as our own. In a real environment there are multiple sources of information as well as multiple ways to extract different channels out of the same data source. Consider a visual colour channel. It can be subdivided into a greyscale video, a colour video, a filtered grayscale video (edges), and many other interesting channels. Besides, there are many more sources of information about the environment such as sound, tactile information, proprioception, and so on.
Suppose one has the capability of implementing a huge number of intentional modules. Suppose there are M incoming sources of information corresponding to as many vectors. For instance, vision could give rise to many different source of information. Sound capability could add a few more sources (different bandwidths, temporal vectors, spectral vectors), and so on. Suppose one has M intentional modules taking care of each of these sources. Suppose that each intentional module has a reasonably large number of elementary units inside. At this point, a few fixed rules will suffice to build dynamically an architecture dubbed here *intentional architecture*.
First, assign to each source of data a separate intentional module. Whether the capacity of the module is saturated (all elementary units are assigned), assign other intentional modules as needed. These modules make the first level of the architecture.
When the first level is complete, use the output of the first level modules as inputs for further levels of intentional modules.
Further levels of intentional modules are assigned to every possible earlier level module's output. However, due to many factors (the richness of the original external source of data, the implemented similarity function inside the elementary units, the incoming data, and so on) the output vector sizes are going to diminish as the level increases. When this happens the intentional module will recruit a smaller and smaller number of elementary units. In that case, its output will get merged with that of other intentional modules with similarly smaller output. Eventually, the previous conditions will obtain for all intentional modules of the higher levels, thereby pushing towards a convergence.
All of the above applies for input and output vectors. As to the control signals the rule is the opposite: backward connecting the higher level intentional modules with the lower level ones. In this way the relevant signal produced by the highest possible elementary units will orient the activity of the lowest level elementary unit modules.

3 A ROBOTIC IMPLEMENTATION

Here a robotic setup, based on a standard commercial robot NAO (Aldebaran Robotics™, Figure 2) with 21 degrees of freedom is taken into consideration. The goal of the setup is to check whether the above-mentioned recipe for generating a cognitive architecture is efficacious in a real environment. Efficiency is not an issue here. The goal is to check whether a whole architecture may be generated by unconstrained interaction with the environment.

The NAO is wirelessly controlled by a software implementation in Visual C#.NET running a multilayered array of such modules on a Intel-based matrix of PCs. Although the robot's basic movements (walking ahead, rotating the body, flexing limbs, closing hands, turning the head, switching on and off the several LEDs) are based on preloaded factory settings, the resulting behaviour is driven by the developing growing network of distributed modules whose structure is fleshed out by the actual coupling between the agent and the environment.

Figure 2. The robot NAO used for this implementation as sold by Aldebaran Robotics™.

Given the rich endowment of sensors, in the presented implementation, only a few of them are connected to the developing cognitive architecture. As is shown in Table 1, the main sensory inputs are connected each to a dedicated intentional module (sometimes grouping together more than one sensory channel). Similarly, each group of basic actions (for instance a simplified repertoire of speech utterances or a simplified repertoire of motor patterns) is connected to other modules. At the beginning, the architecture is empty and each of its modules is empty. Then the system is switched on. Each module begins to receive new data and, according to a set of hardwired bootstrapping criteria, begins to burn the elementary units thereby beginning to get coupled with the particular environment.

Sensor	Intentional Module (level 0)	Intentional Module (further level)	Motor Module	Basic actions
32 x Hall effect sensors + 2 x bumpers	1(A)	?	1(L)	Head 2 DOF Arm 2x5 DOF Pelvis 1 DOF Leg 2x5 DOF Hand 2x1 DOF
1 x accel. 3 axis + 1 x gyro. 2 axis	2(B)	?		
2 x I/R	Unallocated	?	1(M)	51 led variously distributed
Tactile sensor	2 (C+D)	?	1(N)	Simplified repertoire of stereotyped
2 CMOS cameras	3 (D, E, F, G)	?		speech utterances
4 Microphones	4 (H, I , J)	?		

Table 1. Tentative allocation of early modules based on the NAO sensor and motor skills

The experiment has been arbitrarily divided into three steps. I stress that such separation is mainly due to practical constraints. Ideally, the following steps or stages may be seamlessly merged together. Currently, it is easier to keep them neatly separate.

The first stage endorses the coupling between the modules of the first layer. In order to bootstrap the sensor and the motor part, the motor parts are randomly activated for a limited period of time. In this way the architecture begins to embody motor patterns (from L to N). At the same time, the architecture begins to receive data from the sensors. It is thus advisable to envisage a period of stimulation of the robot (showing it objects, helping it to move in the lab, and the like). Such incoming patterns begin to fill the intentional modules assigned to them (from A to J). Each module fills accordingly to the richness of the corresponding channel. If, due to some contingent and unexpected factor, that sensor channel provides a poor input, the allocation of elementary units will be consequently poor. For instance, if the NAO's cameras were bandaged, they would not receive any data and the corresponding intentional units would under-develop.

For each intentional module a maximum number of available elementary units is fixed (in the current implementation it is arbitrarily fixed at 1K). This means that a very rich period of development is going to consume all the available resources.

Either when the intentional modules of the first level (level 0) have saturated their capacities or when a arbitrarily length of time has elapsed, the architecture is ready to enter into a second stage. The architecture must begin to develop by allocating new modules that integrate the output of the modules in the first level and the output of the motor modules and the sensory modules. There are two possibilities: the allocation is either done dynamically on the basis of rules of thumb (at this stage) or is hardwired. At present, it is done using various heuristics. One possible way is to consider all possibilities (as shown in Table 2) and then to check, after a given amount of time, which combinations are the more useful ones. How? By measuring which modules are more active and thus providing the richer flow of data.

		Contact	Posture	Sound	Vision
		A+C+D	B	H+I+J	D+E+F+G
Contact	A				
Posture	B	1			
Sound	H+I+J	2	-		
Vision	D+E+F+G	2	2	3	

Table 2. Tentative allocation of second layer modules. The value in the matrix stands for the number of higher level modules.

The same kind of dynamic allocation is thus performed in the third stage between motor patterns and higher level sensory modules. The objective is to single out sensory-motor patterns which are coupled with the actual environment. In Table 2 is shown an average outcome of such a development. The word 'average' means that different environments may produce a totally different allocation of resources since the sensory channel may receive more or less rich sets of stimuli. The results in Table 2, therefore, are to be taken as an example. In that particular example, there were many visual stimuli associated with sounds.

A similar matrix of associations is finally allocated between the motor patterns and the higher level sensory patterns. Hopefully, the architecture will be capable of picking out sensory-motor contingencies of some relevance in the robot's environment.
Finally, it must be mentioned that the relevant signals of each module provide the control signals that tune the overall behaviour of the developing architecture. These signals acts both as bottom-up and as top-down controls.

4 CONCLUSION AND FUTURE WORK

Right now, the setup is still under continuous development. The objective is to design an architecture that will develop autonomously, getting more and more coupled with a specific environment. At present, the architecture develops and generates coupled patterns at various levels. A further important step will consist in exploiting the capacity, embedded in the intentional modules, to single out new goals to achieve by the architecture. If this is obtained, the architecture may begin to pursue goals and objectives of its own [15].
Moreover, the resulting mesh of modules and units is highly integrated together and it will be interesting to compare the resulting data mesh with other architectures by using indexes of data integration such as liveliness [27-28].
More refined implementations are, of course, to be expected in order to verify the capability of the architecture to foster a real behaviour and to create a complex hierarchy of integrate sensory-motor patterns tightly coupled with the surrounding environment.

The proposed architecture envisages the implementation of a very strong criterion for situatedness. As I mentioned at the start of this paper – *an agent is situated in a given environment to the extent that its cognitive structures are the result of developing inside that environment*. This is what the presented architecture does, notwithstanding all its current shortcomings. Apart from a few hardwired bootstrapping criteria, which the developing modules eventually overcome, the architecture development is totally driven by environment and by the incoming stimuli.
Finally, I would like to address why externalism might endorse phenomenal experience. To do so, I will take advantage of a proposal I presented elsewhere [13, 15, 29] – namely that a perceptual phenomenal experience might be nothing but the external object tightly coupled to the agent's body by means of the cognitive development. In a nutshell, the phenomenal experience of *X* might be nothing but the fact that *X* plays the twofold role of the cause of development *and* a current cause of behaviour. The experience of something would be literally constituted by that something. Consciousness would then be situated in the environment in a very strong sense. This causal condition might endorse, clearly, an externalist model of consciousness offering an interesting conceptual and theoretical framework for machine consciousness since it suggests that consciousness is a matter of the right kind of causal entanglement with the environment.

Acknowledgements. This work has been possible thanks to the support by the Italian-Korea bilateral project between ICT-CNR and KAIST.

REFERENCES

[1] R. Menary. (Ed.) The Extended Mind. MIT Press, Cambridge (Mass) (2010).
[2] M. Anderson. Embodied cognition: A field guide. Artificial Intelligence, 149: p. 91-130 (2003).
[3] S. Harnad. Grounding symbolic capacity in robotic capacity, In: "Artificial Route" to "Artificial Intelligence": Building Situated Embodied Agents, L. Steels and R.A. Brooks, (Eds.), Erlbaum, New York (1995).
[4] M. Shanahan. Embodiment and the Inner Life. Cognition and Consciousness in the Space of Possible Minds. Oxford, Oxford University Press (2010).
[5] A. Clark. Perception, action, and experience: unraveling the golden braid. Neuropsychologia, 47: p. 1460-8 (2009).
[6] R. Chrisley and T. Ziemke. Embodiment, In: Encyclopedia of Cognitive Science, Macmillan, London (2002).
[7] R.A. Wilson. Boundaries of the Mind. The Individual in the Fragile Sciences. Cambridge (Mass), Cambridge University Press (2004).
[8] P. Robbins and M. Aydede. (Eds.) The Cambridge Handbook of Situated Cognition. Cambridge University Press, Cambridge (2009).
[9] F. Adams and K. Aizawa. The bounds of cognition. Philosophical Psychology, 14: p. 43-64 (2001).
[10] R.D. Rupert. Challenges to the Hypothesis of Extended Cognition. The Journal of Philosophy, 101: p. 389-428 (2004).
[11] S. Torrance. In search of the enactive: Introduction special issue on enactive experience. Phenomenology and the Cognitive Sciences, 4: p. 357-368 (2006).
[12] E. Thompson and F.J. Varela. Radical embodiment: neural dynamics and consciousness. Trends in Cognitive Sciences, 5: p. 418-425 (2001).
[13] R. Manzotti. A Process Oriented View of Conscious Perception. Journal of Consciousness Studies, 13: p. 7-41 (2006).
[14] R. Manzotti. No Time, No Wholes: A Temporal and Causal-Oriented Approach to the Ontology of Wholes. Axiomathes, 19: p. 193-214 (2009).
[15] R. Manzotti. A Process-oriented Framework for Goals and Motivations in Biological and Artificial Agents, In: Causality and Motivation, R. Poli, (Ed.), Ontos-Verlag, Frankfurt. p. 105-134 (2010).
[16] R.G. Millikan. Content and vehicle, In: Spatial Representation, N. Eilan, R. McCarthy, and B. Brewer, (Eds.), Blackwell, Oxford (1993).
[17] S. Hurley. The Varieties of Externalism, In: The Extended Mind, R. Menary, (Ed.), MIT Press, Cambridge (Mass). p. 101-155 (2010).
[18] F. Dretske. Knowledge & the flow of information. Cambridge (Mass), MIT Press. xiv, 273 (1981).
[19] S. Harnad. The Symbol Grounding Problem. Physica, D: p. 335-346 (1990).
[20] H. Putnam. Mind, language, and reality. Cambridge, Cambridge University Press. xvii, 457 (1975).
[21] R. Manzotti. A process based architecture for an artificial conscious being, In: Process theories, J. Seibt, (Ed.), Kluwer Academic Press, Dordrecht. p. 285-312 (2003).
[22] R. Manzotti and V. Tagliasco. From "behaviour-based" robots to "motivations-based" robots. Robotics and Autonomous Systems, 51(2-3): p. 175-190 (2005).
[23] A. Chella and R. Manzotti. Artificial Consciousness, In: Perception-Action Cycle: Models, Architectures, and Hardware, V. Cutsuridis, A. Hussain, and J.G. Taylor, (Eds.), Springer, Dordrecht. p. 637-671 (2011).
[24] R. Manzotti and V. Tagliasco. From behaviour-based robots to motivation-based robots. Robotics and Autonomous Systems, 51: p. 175-190 (2005).
[25] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern classification. New York, Wiley. xx, 654 (2001).
[26] R. Manzotti. Towards Artificial Consciousness. APA Newsletter on Philosophy and Computers, 07(1): p. 12-15 (2007).

[27] I. Aleksander and D. Gamez. Informational Theories of Consciousness: A Review and Extension. In: BICS 2010 Conference on Brain-Inspired Cognitive Systems. Madrid (2010).
[28] G. Tononi. An information integration theory of consciousness. BMC Neuroscience, 5: p. 1-22 (2004).
[29] A. Chella and R. Manzotti. Machine Consciousness: A Manifesto for Robotics. International Journal of Machine Consciousness, 1: p. 33-51 (2009).

A Cognitive Neuroscience-inspired Codelet-based Cognitive Architecture for the Control of Artificial Creatures with Incremental Levels of Machine Consciousness

Klaus Raizer, André L. O. Paraense and Ricardo R. Gudwin[1]

Abstract. The advantages given by machine consciousness to the control of software agents were reported to be very appealing. The main goal of this work is to develop artificial creatures, controlled by cognitive architectures, with different levels of machine consciousness. To fulfil this goal, we propose the application of cognitive neuroscience concepts to incrementally develop a cognitive architecture following the evolutionary steps taken by the animal brain. The triune brain theory proposed by MacLean and also Arrabale's ConsScale will serve as roadmaps to achieve each developmental stage, while iCub - a humanoid robot and its simulator - will serve as a platform for the experiments. A completely codelet-based system "Core" has been implemented, serving the whole architecture.

1 INTRODUCTION

1.1 Motivation

In this work, we are particularly interested in studying the cognitive architectures which were proposed to deal with the issue of consciousness [10, 24, 37]. Our main goal is to develop artificial creatures with different levels of machine consciousness, controlled by such architectures. To fulfil this goal, we propose the application of cognitive neuroscience concepts to incrementally develop a cognitive architecture following the evolutionary steps taken by the animal brain.

Looking for inspiration in nature has been a successful way of discovering new solutions for problems in the fields of control, optimization, classification and artificial intelligence. Machine learning techniques such as genetic algorithms, ant colony optimization and neural networks are some examples of the remarkable muse nature can be [27, 25, 36, 16].

The advantages given by machine consciousness have been reported to be very appealing [20, 10, 9].

Nevertheless, the cognitive architectures which are able to benefit from it are not so many, and still under heavy development. So, the motivation to propose and implement yet another cognitive architecture, when there are so many of them already available, lies in the need for an architecture coherent with our hypothesis of a conscious codelet-based artificial mind, able to implement the animal brain in its different evolutionary steps, and in the search for the sufficient feature set in the architecture for each of those steps that matches the results of natural selection along history.

1 School of Electrical and Computer Engineering (FEEC), University of Campinas (UNICAMP), Brazil, email: gudwin@dca.fee.unicamp.br

1.2 Statement and Background of Research

An artificial creature is an autonomous agent, a system embedded in an environment, sensing and acting on it, over time, in pursuit of its own agenda [21]. It can be controlled by a cognitive architecture, which includes aspects of the creature such as memory and functional processes [29], providing a framework to support mechanisms for perception, action, adaptation and motivation [39].

Cognitive architectures are control systems architectures inspired by scientific theories developed to explain cognition in animals and in men. These architectures are typically organized in layers [2, 13], with each layer representing a different level of control and specialized modules [15]. The most famous general cognitive architectures are SOAR [28] and ACT-R [1]. More recently, many specialized cognitive architectures have been proposed, emphasizing different aspects of cognition, e.g. emotions, attention, memory, consciousness and language. Each one has advantages and shortcomings, when compared to each other.

Looking for inspiration in cognitive neuroscience, current research on artificial creatures has focused on the implementation of machine consciousness, with one of its major functions being to recruit relevant resources for solving new or difficult problems [33]. Recently, the study of machine consciousness in cognitive architectures applied to artificial creatures has particularly been exploited [4, 15, 33, 8].

Even though there is not a consensus on what exactly is meant by "machine consciousness", as different authors indeed have different perspectives on what they mean by "consciousness", in a previous work from our group [35], we investigated one interesting proposal, called the Baars-Franklin architecture. During this investigation, we evaluated the possible benefits that such "consciousness" technology, when applied to the control of autonomous agents, could bring to such systems. Our main findings were that the main benefits brought by consciousness (as defined in Baars-Franklin architecture), are:

- Executive Summary of Perception
- The Possibility of Behaviour Automatization

These two advantages arose from the main perspective on what is consciousness after all in the Baars-Franklin architecture, following Dennett [14]. According to this perspective,

consciousness is the emergence of a serial stream on top of a parallel set of interacting devices. In the Baars-Franklin architectures, such devices are called “codelets” [33] (following Hofstadter [26]), which are small pieces of code - similar to Ornstein’s small minds, Minsky’s agents, Edelman’s neuronal groups and Jackson’s demons [19] - specialized in performing simple tasks [33]. This serial stream evidences the most important information flowing in the parallel system at each time instant, creating what we called the executive summary of perception, a special kind of attention mechanism. This serial stream is then broadcast to all codelets in the system, allowing decision-making on both up-to-date input data and filtered relevant information at each time instant. The possibility of automatizing behaviour is also an emergent offspring of this serial stream.

Unconscious behaviour is usually automatic reactive behaviour, performed in parallel by the system codelets. The serial stream can be used then to learn such automatic behaviour, by performing a deliberative one, which is further automatized, giving rise to future automatic behaviours. With this, conscious systems do have an interesting hybrid reactive-deliberative kind of learning, in which new capabilities can be acquired, enhancing the overall behaviour of the system.

Cognitive architectures tend to model functions performed by structures of the animal brain. These structures have, however, changed over millions of years of evolution. One model used to explain this process was the triune brain concept proposed by MacLean [31, 30], which states that the brain developed into a three-layered organ: reptilian brain, paleomammalian brain and neomammalian brain, as can be seen in Figure 1.

Figure 1. The triune brain model as proposed by MacLean. Source: MacLean 1990 [31]

The reptilian brain is composed by the oldest structures that dominate in the brains of snakes and lizards, with a major role on fixed, instinctual behaviours and control for survival. The paleomammalian brain is layered over the reptilian brain, with a major role in emotions (emotional valence and salience) and is better at learning from experience. Finally, the most recent layer is the neomammalian brain, which is the home of complex cognition, deliberative planning, social abilities and language. However controversial this separation in three distinct layers might be today, it remains a helpful way to think about the mammalian brain [11], especially in a computational sense.

MacLean himself points out to the fact that, despite their capacity for operating independently to a certain level, the triune brain is not just a consecutive layering of these three neural structures but actually an integration between the information they share and produce, so “the whole is greater than the sum of its parts". There is often no consensus about which brain structures compose each layer, and the functions each particular structure performs are not easy to discriminate – due, among other reasons, to how massively interconnected most parts of the brain are. As Fuster exemplifies in his book [22], attributing a particular movement control to a given area, or speech only to Broca’s area, ignores the fact that both functions depend on many other neural structures. With that in mind, this work aims at avoiding direct mappings [34] between neural structures and its functions, and focuses instead on a framework developed over a large body of brain and psychological evidence. The proposed architecture is based on Baars and Gage’s functional framework, as seen in Figure 2, to develop a codelet-based, biologically plausible cognitive architecture.

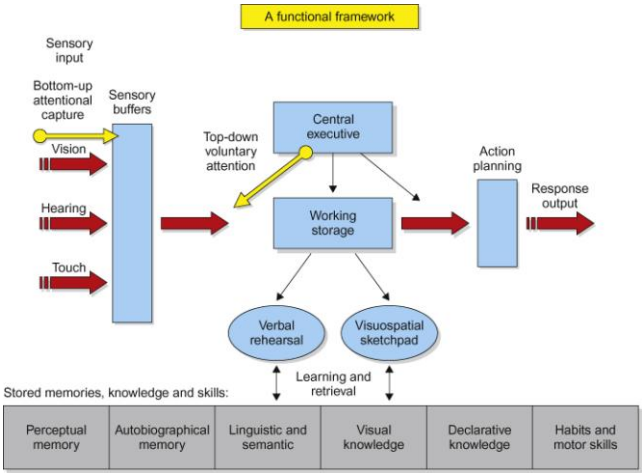

Figure 2. Baars and Gage’s functional framework. Source: Baars and Gage 2010 [11], with kind permission.

In the framework from Figure 2 each sense has a brief storage ability, also called sensory buffer. Elements in the sensory buffer are modified by bottom-up selective attention, which happens in vision for instance when confronting particular patterns, or in hearing when there is a loud noise. There is a top-down component to selective attention coming from the central executive, which allows voluntary attention to happen. The central executive is part of working memory, as defined by Baddeley [12], and it exerts supervisory control over all voluntary activities. Working storage is a short term and dynamic storage mechanism. It is composed of active populations of neurons which can consolidate into long term memories and is believed to have very limited capacity. The verbal rehearsal and the visuospatial sketchpad involve mental capacities used to remember things like new words (in the case of inner speech), faces, or spatial information (in the case of visuospatial sketchpad). They are both linked to long-term memory by a learning and retrieval mechanism. Long-term memory is represented by the gray boxes on the bottom and is

comprised by a number of different types of memory, each with its own functions and characteristics. At the right side of the diagram, there is action planning, which can have both conscious and unconscious components, and finally an output response that closes the perception-action cycle.

2 METHODOLOGY

The work is following a path similar to the evolutionary steps taken by the animal brain as stated by MacLean [31]. The hypothesis is that such an approach should guarantee a grounded intelligent system at each developmental phase, while biasing the system towards high-level animal intelligence.[2]

The initial development of this architecture emphasizes on behavioral results. A low level approach to evaluate its consciousness levels - such as information integration, as seen in the works of Tononi [38] - might be implemented in future works. ConsScale, a biologically inspired scale designed to evaluate the presence of cognitive functions associated with consciousness [5, 3, 6], will be used to assess the different levels of control implemented within this cognitive architecture. There are two ways of using ConsScale: the Standard Evaluation Process (SEP) and the Simplified Rating Process (SRP). SEP is used to evaluate existing implemented agents, providing an accurate measure of the agent's cognitive level. SRP, on the other hand, is used as an approximation of the potential level of an existing or still to be implemented model.

In this initial work, the SRP for each level of development - reptilian, paleomammalian, neomammalian and *Homo sapiens* - is calculated, and a roadmap of behaviour profiles (BP) is proposed in order to reach an accurate measure of a specific domain. The ConsScale Quantitative Score (CQS), spanning from 0 to 1000 in an exponential fashion, is also calculated for each stage, providing a numerical value indicating their cognitive power.

The platform used as an specific domain for the initial experiments is the iCub humanoid robot simulator [32], but the architecture is built so it can be applied to different platforms and applications. This platform was chosen because it provides a "Human-Like" architectural level, as described by Arrabales [5, 6].

2.1 Conscious Codelet-Based Cognitive Architecture

Figure 3 shows an UML class diagram describing the architecture's modules.

The relationship between modules and their features according to the functional framework of Figure 2 is better understood by following a full cognitive cycle, considering the neomammalian brain:

1. Sensors (BodyInterface) get information from the World (iCub) and send it to the Sensory Buffer (BodyInterface);
2. Bottom-up Attention (Perception) acts on Sensory Buffer (BodyInterface), giving rise to objects from raw sensory inputs;

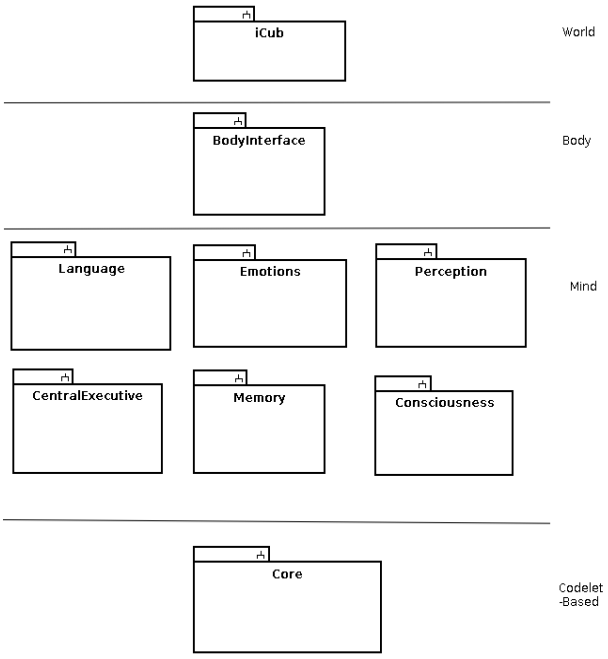

Figure 3. Architecture's layers and subsystems

3. Bottom-up Attention (Language) acts on Sensory Buffer (Perception), giving rise to symbols from objects;
4. Top-down Attention (Central Executive) acts on Sensory Buffer's objects (Perception) and symbols (Language);
5. Top-down Attention (Central Executive) brings information into Working Storage (Memory);
6. Learning and Retrieval Mechanism (Memory) consolidates to Stored Memory (Memory) and brings into Working Storage (Memory) long-term information;
7. Spotlight Controller (Consciousness) acts on Working Storage (Memory), defining Spotlight (Consciousness) content;
8. Action Selection (Emotions) uses information under Spotlight (Consciousness) to select a plan, composed by a list of behaviours;
9. Action Selection (Central Executive) uses information under Spotlight (Consciousness) to select a plan, composed by a list of behaviours;
10. Behaviour sequence is sent to Action Buffer (BodyInterface);
11. Actuators (BodyInterface) act on the World (iCub) based on Action Buffer (BodyInterface).

Figure 4 shows a diagram depicting how the concepts of the codelet-based Core subsystem have been implemented. Following this picture, Memory Objects are single units of data in memory, which have a type (T) and some information (I). The

[2] It is important to acknowledge, however, that the path taken by mammals in evolution, especially the case of *Homo sapiens*, is not the only one that led to high-level cognition. Examples of high level cognitive behaviour, and potentially conscious capabilities, have been observed in modern birds and cephalopods [17].

Raw Memory contains all Memory Objects in the system. It can be logically divided in different kinds of memory, such as the stored memories from Figure 2. Codelets are devices which are composed by small pieces of code, specialized in performing simple tasks (proc), a list of input Memory Objects (In), the ones that are read, a list of output Memory Objects (Out), the ones that are written, a list input broadcasted Memory Objects (B), the ones that were broadcasted by consciousness mechanisms, and an activation level (A). Coalitions are groups of Codelets which are gathered in order to perform a task by summing up their abilities. Two or more Codelets share a Coalition when they write in and/or read from the same set of Memory Objects. The Coderack (following Hofstadter [26]) is the pool of all active Codelets in the system.

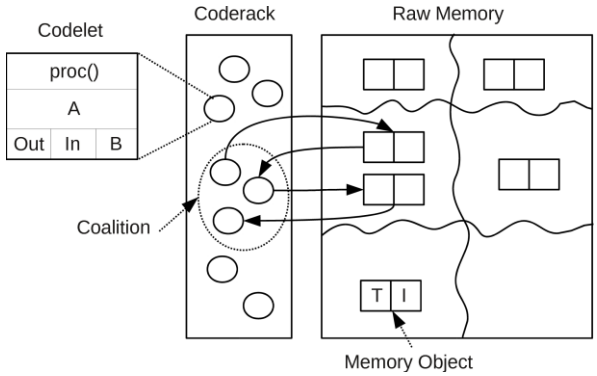

Figure 4. Core's concepts

The other modules are the subject of future work and will be implemented according to the planned steps of the architecture, as further explained in Section 2.2.

2.2 Evolutionary Steps Taken by the Animal Brain

2.2.1 Reptilian

According to MacLean, the protoreptilian formation is composed by a group of ganglionic structures located at the base of the forebrain of reptiles, birds and mammals. It is also known as the striatal complex and brainstem [11] or, as he puts it, the R-Complex. It was traditionally thought to be the motor apparatus under control of motor cortex and reveals a number of fixed behaviours (25 special forms of behaviours and 6 forms of what he calls "interoperative" behaviours [31]), involved in the regulation of the animal's daily routines. In this sense, the R-Complex is composed by pre-programmed regulators for homeostasis and survival, lacking an advanced learning mechanism as the one seen latter in evolution. In the one hand, it excels at performing sensory categorization, such as identifying a particular smell as being harmful or not, and then generating reflexive messages about what to do, like running or biting [23]. On the other hand, it is constrained by its daily master routine, destined to perform a limited number of behaviours. Reptiles and lizards however need to "learn" their territory in order to know which hole to escape into in case of a predator or just to find their way home. This, and other examples of very basic "memory/learning" mechanisms, is what MacLean called protomentation, which are "rudimentary mental processes that underlie a meaningful sequential expression of prototypical patterns of behaviour" [31].

Based on the main characteristics of a creature with an R-Complex, a number of skills from ConsScale are proposed for this level of development:

CS2;1 : Fixed reactive responses.

BP2;1 : Basic reflexes such as blinking and contraction of limbs as responses to pain.

CS3;3-5 : Selection of relevant sensory/memory/motor information.

BP3;3-5 : The robot reacts to predefined sensory inputs and stores basic information in fixed memory.

CS3;6 : Evaluation (positive or negative) of selected objects or events.

BP3;6 : The robot evaluates sensory input comparing it with predefined patterns to evaluate sensory inputs as being good or bad for it.

CS4;2 : Directed behaviour toward specific targets like following or escape.

BP4;2 : The robot selects grabbing action towards regions that seem good for it.

According to ConsScale, this selection of skills constitutes a level 2 (Reactive) agent, with a CQS of 0.21 in a scale from 0 to 1000. This set of skills suggest the need for a Body Interface module, responsible for dealing with all somato-sensory data exchange - sensory input and response output from Figure 2 - between agent and environment. This module also holds a fixed number of reactive responses and autonomic behaviours that have as a primary concern the survival and self-preservation of the agent. Perception at this level is very basic, with low resolution pattern recognition and a bottom-up attention mechanism that depends on the agent's nature and objectives. The output functions are organized in the Central Executive module, which at this level of development is responsible for action selection with a repertory of fixed predefined behaviours. The agent at this level lacks a general Memory System, counting only on predefined memory slots for performing specific tasks.

There is a great debate on whether creatures other than humans do or do not possess consciousness as we experience it. One major problem faced by such a debate is the lack of an accurate definition of what consciousness is and what is needed for its emergence. In this work, machine consciousness is the implementation of Global Workspace (GW) theory [7] as a means to achieve primary consciousness, in which percepts are united into episodic scenes [18]. In this sense, it is assumed here that a protoreptilian brain lacks the reentrant interactions in the thalamocortical system needed to sustain consciousness.

2.2.2 Paleomammalian

The next evolutionary step in the development of the vertebrate brain is the paleomammalian formation. With this new set of neuronal structures - some notable examples being the amygdala, hippocampus and hypothalamus - also known as the limbic system, animals became able to experience emotions [11], which are essentially the capacity of turning up or down the "volume" of drives that guide behaviour for survival. This emotional "skill" greatly affects and communicates with the

aforementioned autonomic system, provoking marked physiological changes within the organism. Learning and memory have also shown remarkable improvement. An animal with a limbic system mounted on top of its R-Complex was able to discriminate good things from bad ones also by looking into its past memories [23]. MacLean emphasized that this part of the mammalian brain is responsible for a number of behaviours and characteristics that were absent in ancient reptiles such as nursing, audio-vocal communication for maternal contact and play [31].

The ConsScale skills added at this level are listed as follows:

CS3;1 : Autonomous acquisition of new adaptive reactive responses.

BP3;1 : Learns to "eat" certain kinds of "food" and reject others.

CS3;2 : Usage of proprioceptive sensing for embodied adaptive responses.

BP3;2 : Looks for "food" when hunger state reaches a certain level. Plays to get happier.

CS3;7 : Selection of what needs to be stored in memory.

BP3;7 : Emotional valence influences selection of what is relevant to be stored in memory.

CS4;1 : Trial and error learning. Re-evaluation of selected objects or events.

BP4;1 : The robot learns what is good (good food) or bad (rotten food) for him by trial and error.

CS4;3 : Evaluation of the performance in the achievement of a single goal.

BP4;3 : Evaluates how successful it is in pursuing a single goal, such as looking for "food" and uses this information to get better at it.

CS4;5 : Ability to build depictive representations of percepts for each available sensory modality.

BP4;5 : The robot can discern particular objects and some of its properties and calculate their relative positions.

CS5;1 : Ability to move back and forth between multiple tasks.

BP5;1 : An interrupted behaviour is resumed later if still relevant. For example: playing with certain objects in the environment can be resumed after stopping this behaviour to satiate hunger.

CS5;4 : Autonomous reinforcement learning (emotional learning).

BP5;4 : The robot calculates a "reward" based on how good it is at a task and improves its performance. It might throw a ball at a given target and get better at it by practicing.

CS5;2 : Seeking of multiple goals

BP5;2 : Having more than one goal, such as satisfying hunger and play, it uses CS5;1 to alternate between them.

CS5;3 : Evaluation of the performance in the achievement of multiple goals.

BP5;3 : The robot evaluates its own performance at pursuing multiple goals, and alternating among them, instead of pursuing only one.

CS5;6 : Ability to generate selected mental content with grounded meaning integrating different modalities into differentiated explicit percepts.

BP5;6 : The contents of the conscious broadcast, defined by the consciousness module, constitute mental content with grounded meaning and it is composed by an integration of percepts from different modalities.

CS6;1 : Self-status assessment (background emotions).

BP6;1 : Evaluates its own inner physical and emotional state and has its global behaviour influenced by it.

CS6;2 : Background emotions cause effects in agent's body.

BP6;2 : The emotional state is reflected into the robot's body (happy or sad faces) through its autonomic functions.

CS6;3 : Representation of the effect of emotions in organism and planning (feelings).

BP6;3 : Together with BP6;1, if the robot is high on health and hungry it may go look for food but if low on health and hungry it might hide at home.

This set of skills appears at the ConsScale as a level 3 (adaptive) agent, with a CQS of 7.21 in a scale from 0 to 1000. Even having a number of higher skills, such as CS6;1(Self-status assessment) for instance, it lacks some dependencies such as CS4;4 (Basic planning capability) that would allow it to attain a higher score.

There is an evolution in perception at this stage so the agent is able to perform higher-level pattern recognition, and discern particular objects in the environment. The central executive performs top-down attention over percepts, providing full attention selection capability. The memory module becomes generic, in the sense that memory objects are produced, stored and retrieved by means of a learning/retrieval mechanism. Those memories and percepts are marked with emotional content, influencing the aforementioned learning/retrieval mechanism.

At this point it is assumed that the reentrant interactions between parts of the thalamocortical system mediating perceptual categorization and those mediating memory have evolved in a way that allows the emergence of primary consciousness. The consciousness module implements GW theory, producing an attentional spotlight that broadcasts its contents to all the system. However, the creature still lacks the capacity to report its conscious stream, an ability human beings possess and which is used in our case to verify the existence of consciousness as we perceive it.

2.2.3 Neomammalian

The neomammalian formation is the latest addition to the vertebrate animal brain. Its distinguished structure is the neocortex which is composed of many layers, with a smooth surface in small mammals and deeply grooved in larger ones. The neocortex is highly oriented toward the external world [31]. With it, animals are capable not only to understand their senses but also to develop a symbolic representation of those senses and inner representations. Being on top of the limbic system, it is also capable of developing feelings about these symbols and abstract ideas [23]. Its most distinctive role, however, lies in what is called executive functions, which consist, among other things, on the ability to organize sequences of actions towards a given goal.

As the vertebrate brain evolved, the organism's actions became more based on its memories and prior experiences than on reflexive responses to the environment based on its needs (as can be seen in the transition from protoreptilian to paleomammalian). These actions also became more deliberate and voluntary [22], especially in the transition from the paleomammalian to the neomammalian brain. With this evolution, important parts of the neocortex such as the prefrontal cortex show significant growth in proportion to more ancient brain structures, with a maximum size achieved only in the human primate [22].

The ConsScale skills at the neomammalian level are as follows:

CS4;4 : Basic planning capability: calculation of next n sequential actions.

BP4;4 : Plans a sequence of actions to attain a goal. Example: playing for learning wastes energy, so it plans a break time to replenish it.

CS5;5 : Advanced planning capability considering all active goals.

BP5;5 : Takes into account information from CS5;3 to improve seeking multiple goals. Such as reducing transition time between behaviours or deciding a better order of behaviours.

CS6;4 : Ability to hold a precise and updated map of body schema.

BP6;4 : It has a map of its own body and can use it to plan/select behaviours.

CS6;5 : Abstract learning (lessons learned generalization).

BP6;5 : Its memories influences how it behaves in a general way. Differently from how it would behave without them.

CS6;6 : Ability to represent a flow of integrated percepts including self-status.

BP6;6 : The consciousness module allows an executive summary, composed by integrated percepts and allowing the robot to represent its self-status.

CS7;1-3 : Representation of the relation between self and perception/action/feelings.

BP7;1-3 : Special codelets are specialized in establishing the relation between perception and action, and the robot's sense of self as an emotional agent.

CS7;4 : Self-recognition capability.

BP7;4 : The robot recognizes itself as an agent in the world, allowing CS7;5.

CS7;5 : Advanced planning including the self as an actor in the plans.

BP7;5 : It performs CS5;5 (advanced planning) taking into account itself as an agent.

CS7;6 : Use of imaginational states in planning.

BP7;6 : The robot estimates future emotional state for possible outcomes due to planned actions and uses this information to select behaviour.

CS7;7 : Learning of tool usage.

BP7;7 : Learns to use objects in the scene to perform tasks, such as throwing a ball at something out of reach to bring it down.

CS7;8 : Ability to represent and self-report mental content (continuous inner flow of percepts/inner imagery).

BP7;8 : The robot can report its conscious contents.

CS8;1 : Ability to model others as subjective selves.

BP8;1 : It will use its own mental model to predict/estimate another's actions.

CS8;2 : Learning by imitation of a counterpart.

BP8;2 : The robot will learn new behaviours, such as waving or selecting particular objects, by watching a counterpart doing it.

CS8;3 : Ability to collaborate with others in the pursuit of a common goal.

BP8;3 : The robot can form plans including other agents to reach a common goal. Such as pushing boulders or exchanging tools.

CS9;3 : Advanced communication skills (accurate report of mental content as basic inner speech).

BP9;3 : The robot is able to report inner mental state.

A neomammalian agent is registered as being level 7 (self-conscious) and scores 207.63 at the CQS scale.

The central executive at this stage is able to produce new behaviours that are added to the repertory of predefined ones. It becomes able to actually devise plans to achieve its goals. The major add-on feature in perception is the creation of memory objects with symbolic content.

The agent now has a Language module which is responsible for producing an accurate report of its mental content. This allows basic reportability tests of consciousness but high-order consciousness [18] should only be achieved at the *Homo sapiens* stage.

Homo sapiens

The most distinguished part of the human brain is its big frontal lobes. These regions have shown remarkable expansion at the last stage of human evolution and can be regarded as the core machinery for what we understand as being human. The aforementioned prefrontal cortex (PFC) plays a decisive role in both social cognition and advanced planning and problem solving. The ability to recombine and manipulate internal representations, a vital skill for the development of advanced language, and the capacity of holding "images of the future", important for tool-making, are both critically dependant on the PFC [11].

The ConsScale skills at the *Homo sapiens* level are as follows:

CS8;4 : Social planning (planning with socially aware plans).

BP8;4 : The robot devises plans including groups of agents in order to improve the group's conditions as a whole.

CS8;5 : Ability to make new tools.

BP8;5 : The robot can combine objects in the scene to produce a new tool. For instance, bending a wire so it works as a hook.

CS8;6 : Inner imagery is enriched with mental content related to the model of others and the relation between the self and other selves.

BP8;6 : Robot's conscious content integrates mental imagery related to its own model and the models of other agents.

CS9;1 : Ability to develop Machiavellian strategies like lying and cunning.

CS9;1 : The robot is able to estimate another agent's reaction to its actions and use it in its own benefit. For instance, if the robot wants a person to get closer, it might ask this person for "food" even without being hungry.

CS9;2 : Social learning (learning of new Machiavellian strategies).

BP9;2 : The robot can learn new strategies as in CS9;2, not implemented *a priori*.

CS9;4 : Groups are able to develop a culture.

BP9;4 : Groups of robots and other agents can develop their own cultural content and pass it on to other individuals to improve learning.

CS9;5 : Ability to modify and adapt the environment to agent's needs.

BP9;5 : The robot can include the altering of the environment in its plans to reach a goal. Such as moving rocks to form a barrier.

CS10;1 : Accurate verbal report. Advanced linguistic capabilities. Human-like inner speech.

BP10;1 : The robot should be able to develop conversations, with grammar and semantic content.

CS10;2 : Ability to pass the Turing test.

BP10;2 : At this point, the robot should be able to pass a domain specific Turing test.

CS10;3 : Groups are able to develop a civilization and advance culture and technology.

BP10;3 : Groups of robots and other agents should be able to interact in a cultural and social way to develop new tools and knowledge about the environment.

At this stage, the agent reaches level 10 (Human-Like) in the ConsScale, with a CQS of 745.74. Higher levels could only be achieved with structural modifications in the basic architecture to allow several streams of consciousness being managed by the same agent[3].

Structurally, the cognitive architecture at the *Homo sapiens* level is the same as the one described for the neomammalian brain. It has the same modules and the communication between them is virtually identical. The difference lies in the codelets used to perform the new skills necessary to achieve this level of cognition.

3 It is not clear if being able to manage many streams of consciousness in the same body would result in an advantage, as suggested by ConsScale. One can argue based on absence that, in the course of evolution, natural selection would have selected mammals or other classes of animals which had appeared, by mutation mechanisms, with more than one stream of consciousness if this fact resulted in an advantage.

3 DISCUSSION

The architecture proposed here, in its many development stages, aim at managing the agent's attentional resources in order to fulfil its tasks and reach its goals. Distinctively from other architectures, this model commits to a single, uniform notation for encoding knowledge, which are memory objects that hold information for different applications. This has the advantage of simplicity and may support learning and reflection more easily, since they have to operate on a single type of structure.

The codelet approach further enhances the modularity and scalability of the system. Particular codelets can be designed on demand to fulfil a given task and be readily implemented in the architecture without the need of major architectural modifications.

Due to its essentially modular structure, as seen in ACT-R and LIDA, this triune cognitive architecture differs from other well known architectures such as SOAR [29, 15]. A modular structure offers a number of advantages, such as robustness and allowing distributed processing. Moreover, ACT-R and SOAR architectures lack a consciousness mechanism, which would allow an efficient perceptual summary and behaviour automation.

A well known cognitive architecture that implements a consciousness mechanism is LIDA[15] which, as previously mentioned, is also a highly modular architecture - but not completely codelet-based - and strongly based on cognitive neuroscience. It aims, among other things, at being a tool for generating testable hypotheses about human and animal cognition, which might make its use at simpler applications problematic.

The architecture here presented deals with this applicability problem by being decomposable into three distinct architectures, each with a level of complexity more suitable to a particular application. In other words, this work employs a technological approach, by drawing inspiration from neuroscience in order to develop better intelligent artificial systems. Many works of this type also aim at having a contribution toward taking the scientific side of this research forward, hoping to better understand or make important discoveries about biological consciousness by building successively more complex artificial agents with cognitive architecture. This is not the case in this work, which aims at taking advantage of the new findings in science to build better technologies.

4 CONCLUSIONS & FUTURE WORK

This is a work in progress, the next step of which is the implementation of each structure of the first brain layer, associated with the reptilian brain, and the integration of these structures in the architecture. The system's "Core" has been implemented in a completely codelet-based fashion, serving the whole architecture, and the "Consciousness" module is being implemented on the basis of the Baars-Franklin concepts.

We would like to thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for the financial support.

REFERENCES

[1] J.R. Anderson, D. Bothell, M.D. Byrne, S. Douglass, C. Lebiere, Y. Qin, et al. (2004) 'An Integrated Theory of the Mind', *Psychological Review*, **111**(4), 1036–1060.

[2] R. Arrabales, A. Ledezma, and A. Sanchis, 'A Cognitive Approach to Multimodal Attention' (2009), *Journal of Physical Agents*, **3**(1), 53–63.

[3] R. Arrabales, A. Ledezma, and A. Sanchis (2009), 'Assessing and Characterizing the Cognitive Power of Machine Consciousness Implementations', in *AAAI Fall Symposium Series*, (2009).

[4] R. Arrabales, A. Ledezma, and A. Sanchis (2009), 'Towards Conscious-like Behavior in Computer Game Characters', in *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*, pp. 217–224. IEEE.

[5] R. Arrabales, A. Ledezma, and A. Sanchis (2010a), 'ConsScale. a pragmatic scale for measuring the level of consciousness in artificial agents', *Journal of Consciousness Studies*, **17**(3-4), 131–164(34).

[6] R. Arrabales, A. Ledezma, and A. Sanchis (2010b), 'The Cognitive Development of Machine Consciousness Implementations', *International Journal of Machine Consciousness*, **02**(02), 213.

[7] B.J. Baars (1988), *A Cognitive Theory of Consciousness*, Cambridge Univ Press.

[8] B.J. Baars (1996), 'In the Theatre of Consciousness. Global Work Space Theory, A Rigorous Scientific Theory of Consciousness', *Journal of Consciousness Studies*, **4**(1), 292–309.

[9] B.J. Baars and S. Franklin (2003), 'How Conscious Experience and Working Memory Interact', *Trends in Cognitive Sciences*, **7**(4), 166–172.

[10] B.J. Baars and S. Franklin (2009), 'Consciousness is computational: the lida model of global workspace theory', *International Journal of Machine Consciousness*, **01**, 23, (2009).

[11] B.J. Baars and N.M. Gage (2010), *Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience*, Academic Press.

[12] A. Baddeley (2003), 'Working Memory and Language: an Overview', *Journal of communication disorders*, **36**(3), 189–208.

[13] R.A. Brooks (1991), 'Intelligence without representation', *Artificial intelligence*, **47**(1-31-3), 139–159.

[14] Daniel C. Dennett (1991), *Conciousness Explained*, Back Bay Books.

[15] S.K. D'Mello, S. Franklin, U. Ramamurthy, and B.J. Baars (2006), 'A Cognitive Science Based Machine Learning Architecture', in *AAAI 2006 Spring Symposium Series Sponsor: American Association for Artificial Intelligence. Stanford University, Palo Alto, California, USA.*

[16] K. Doerner, W. Gutjahr, R. Hartl, C. Strauss, and C. Stummer (2004), 'Pareto Ant Colony Optimization: A Metaheuristic Approach to Multiobjective Portfolio Selection', *Annals of Operations Research*, **131**, 79–99. 10.1023/B:ANOR. 0000039513.99038.c6.

[17] D.B. Edelman, B.J. Baars, and A.K. Seth (2005), 'Identifying hallmarks of consciousness in non-mammalian species.', *Consciousness and Cognition*, **14**(1), 169–87.

[18] G.M. Edelman, *Wider than the Sky: The Phenomenal Gift of Consciousness*, Yale Univ Pr, 2004.

[19] S. Franklin, *Artificial Minds*, The MIT Press, 1997.

[20] S. Franklin, B.J. Baars, U. Ramamurthy, and M. Ventura, 'The Role of Consciousness in Memory', *Brains, Minds and Media*, **1**(1), 38, (2005).

[21] S. Franklin and A. Graesser, 'Is it an Agent, or just a Program? : A Taxonomy for Autonomous Agents', *Intelligent Agents III Agent Theories, Architectures, and Languages*, 21–35, (1997).

[22] J.M. Fuster, *The Prefrontal Cortex*, Academic Press, 2008.

[23] D. Goleman, *Emotional Intelligence*, Bantam Dell Pub Group, 2006.

[24] P.O. Haikonen (2007), *Robot Brains: Circuits and Systems for Conscious Machines*, Wiley-Interscience.

[25] S. Haykin (1999), *Neural Networks: a Comprehensive Foundation*, Prentice Hall, 2 ed. edn.

[26] D.R. Hofstadter and M. Mitchell (1994), 'The copycat project: A model of mental fluidity and analogy-making.', *In Holyoak, K.J & Barnden, J.A. (Eds.). Advances in connectionist and neural computation theory*, **2**, 31–112.

[27] J.H. Holland (1992), *Adaptation in Natural and Artificial Systems*,Cambridge, MA: MIT Press.

[28] J.E. Laird (2008), 'Extending the Soar Cognitive Architecture', in *Proceeding of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, pp. 224–235, Amsterdam: IOS Press.

[29] P. Langley, J. E. Laird, and S. Rogers (2009), 'Cognitive architectures: Research issues and challenges', *Cognitive Systems Research*, **10**(2), 141–160.

[30] P.D. MacLean (1985), 'Brain Evolution Relating to Family, Play, and the Separation Call', *Archives of General Psychiatry*, **42**(4), 405.

[31] P.D. MacLean (1990), *The Triune Brain in Evolution: Role in Paleocerebral Functions*, NY: Springer.

[32] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori (2008), 'The iCub Humanoid Robot: an Open Platform for Research in Embodied Cognition', in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, pp. 50–56. ACM.

[33] A.S. Negatu and S. Franklin (2000), 'An Action Selection Mechanism for Conscious Software Agents 1', *Cognitive Science Quarterly*, 1–21.

[34] F. Rodrıguez, F. Galvan, F. Ramos, E. Castellanos, G. Garcıa, and P. Covarrubias (2010), 'A Cognitive Architecture Based on Neuroscience for the Control of Virtual 3D Human Creatures', in *Brain Informatics: International Conference, (BI-2010)*, p. 328.

[35] R.C.M. Silva and R.R. Gudwin (2010), 'An Introductory Experiment with a Conscious-Based Autonomous Vehicle', in *Proceedings of ROBOCONTROL 2010 - IV Workshop in Applied Robotics and Automation*, pp. 1–9.

[36] B. Suman and P. Kumar (2005), 'A Survey of Simulated Annealing as a Tool for Single and Multiobjective Optimization', *Journal of Oper Res Soc*, **57**, 1143 – 1160.

[37] R. Sun (2007), 'The Importance of Cognitive Architectures: An Analysis Based on CLARION', *Journal of Experimental & Theoretical Artificial Intelligence*, **19**(2), 159–193.

[38] G. Tononi (2008), 'Consciousness as Integrated Information: A Provisional Manifesto.', *The Biological bulletin*, **215**(3), 216–42.

[39] D. Vernon, G. Metta, and G. Sandini (2007), 'The iCub Cognitive Architecture: Interactive Development in a Humanoid Robot', *2007 IEEE 6th International Conference on Development and Learning*, 122–127.

Self System in a Model of Cognition

Uma Ramamurthy* and Stan Franklin**

* St Jude Children's Research Hospital, Memphis, TN 38105, USA.
Email: **uma.ramamurthy@stjude.org**.
**Dept. of Computer Science and Institute for Intelligent Systems,
The University of Memphis, Memphis, TN 38152, USA.
Email: **franklin@memphis.edu**.

Abstract. Philosophers, psychologists and neuroscientists have proposed various forms of a "self" in humans and animals. All of these selves seem to have a basis in some form of consciousness. The Global Workspace Theory (GWT) [1 - 3] suggests a mostly unconscious, many layered self-system. In this paper we consider several issues that arise from attempts to include a self-system in a software agent/cognitive robot. We explore these issues in the context of the LIDA model [4], [15] which implements the Global Workspace Theory.

1 INTRODUCTION

The LIDA model is both a conceptual and computational model implementing and fleshing out a major portion of Global Workspace Theory (GWT) [1]. The model also implements a number of other psychological and neuropsychological theories including situated cognition [20], perceptual symbol systems [21], working memory [23], memory by affordances [24], long-term working memory [25], Sloman's H-CogAff [26], and transient episodic memory [22].

As is true with any computational/conceptual model of human cognition, the LIDA model has gaps, areas in which it cannot yet offer explanations. One such gap is the self-system.

Baars [1] sees the self as an unconscious executive that receives conscious input and controls voluntary actions. There is a direct connection between self and consciousness. If one damages the self-system of a human, then conscious contents may also disappear. Recall that in people with split brains, the dissociated executive loses access to the conscious contents of the other executive [1], [6]. Our goal is to implement a self-system in the LIDA model that is in tune with GWT, while attempting to understand how the self system works in humans/animals.

2 SELF SYSTEM

In the spirit of GWT, a self-system in an autonomous agent may be constituted by three major components namely, the *Proto-Self*, the *Minimal (Core) Self* and the *Extended Self* as shown in Figure 1.

Neuroscientist Antonio Damasio conceived a *proto-self* as a short-term collection of neural patterns of activity representing the current state of the organism [9]. This proto-self receives neural and hormonal signals from visceral changes.

The *minimal* or *core self* is attributed to all animals by biologists, philosophers and neuroscientists [9], [12], [19]. The core consciousness is continually regenerated in a series of pulses (LIDA's cognitive cycles [11]), which blend together to give rise to a continuous stream of consciousness. The minimal or core self is partitioned into the self-as-agent (the acting self), the self-as-experiencer (the experiencing self) and the self-as-subject (the self that can be acted upon by other entities in the environment).

The *extended self* consists of the autobiographical self, the self-concept, the volitional or executive self, and the narrative self. This extended self is ascribed to humans and, possibly, to higher animals. The autobiographical self develops directly from episodic memory [7], [10]. The self concept, also referred to as the self context [1] or the selfplex [8] consists of enduring self beliefs and intentions, particularly those relating with personal identity and properties. The volitional self provides executive function [1]. Finally, the narrative self is able to report, sometimes equivocally, contradictorily or self-deceptively, on actions, intentions, etc., [13].

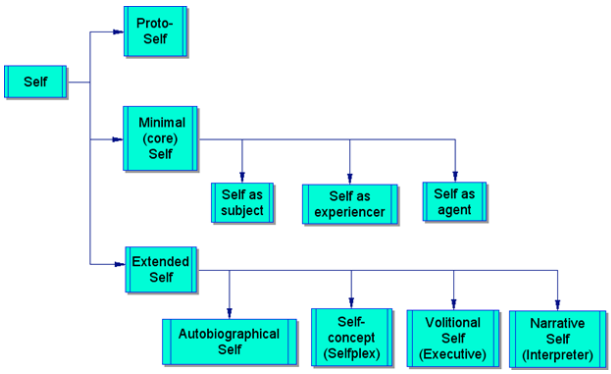

Figure 1. The Self System for LIDA

3 LIDA MODEL

The LIDA computational architecture, derived from the LIDA cognitive model, employs several modules that are designed using computational mechanisms drawn from the "new AI." These include variants of the Copycat Architecture [27], [30], Sparse Distributed Memory [28], the Schema Mechanism [31], [33], the Behavior Net [29], and the Subsumption Architecture [32]. As the architecture implements GWT, the various modules in this system have processors executing and accomplishing small, simple and complex tasks. These processors are often

represented by codelets which are small pieces of code that accomplish one specific task. The LIDA model has been detailed in several publications [34], [35], [36].

LIDA's processing can be viewed as consisting of a continual iteration of Cognitive Cycles [11], [35]. Each cycle constitutes units of understanding, attending and acting. During each cognitive cycle a LIDA-based agent first makes sense of its current situation as best as it can by updating its representation of its world, both external and internal. By a competitive process, as specified by Global Workspace Theory, it then decides what portion of the represented situation is most in need of attention. Broadcasting this portion, the current contents of consciousness, enables the agent to finally choose an appropriate action which it then executes. Thus, the LIDA cognitive cycle can be subdivided into three phases, *the understanding phase, the consciousness phase*, and *the action selection phase*.

Beginning the *understanding phase*, incoming stimuli activate low-level feature detectors in Sensory Memory. The output is sent to Perceptual Associative Memory where higher-level feature detectors feed into more abstract entities such as objects, categories, actions, events, etc. The resulting percept is sent to the Workspace where it cues both Transient Episodic Memory and Declarative Memory producing local associations. These local associations are combined with the percept to generate a current situational model, the agent's understanding of what's going on right now.

Attention Codelets begin the *consciousness phase* by forming coalitions of selected portions of the current situational model and moving them to the Global Workspace. A competition in the Global Workspace then selects the most salient coalition whose contents become the content of consciousness that is broadcast globally.

In the *action selection phase* of LIDA's cognitive cycle, relevant action schemes are recruited from Procedural Memory. A copy of each such is instantiated with its variables bound and sent to Action Selection, where it competes to provide the action selected for this cognitive cycle. The selected instantiated scheme triggers Sensory-Motor Memory to produce a suitable algorithm for the execution of the action. Its execution completes the cognitive cycle.

4 IMPLEMENTING SELF SYSTEM IN LIDA

In the context of the LIDA model briefly described in the previous section, let us consider how the various parts of a Self-System in Figure 1 can be implemented in this model.

Implementing Proto-Self: The Proto-Self for a software agent or cognitive robot can be viewed as the set of global and relevant parameters in the various modules of the autonomous agent. In LIDA, these are the parameters in the Behavior Net, the memory systems, and the underlying computer system's memory and operating system. These aspects which constitute the Proto-Self are already present in the LIDA model.

Implementing Minimal/Core Self: All the three parts of Minimal Self can be implemented as sets of entities in the LIDA ontology, that is, computationally as collections of nodes in the slipnet of LIDA's perceptual associative memory.

One of the features of consciousness is subjectivity, the first person point of view. The self-as-agent accomplishes some aspects of such subjectivity. Self-as-agent can be implemented as the set of self-action nodes in the slipnet, i.e., nodes representing actions by the agent such as lie-down, stand, roll-over, walk, glance-left, etc. Having such action nodes in the slipnet would allow actions –

- to be part of structure building in working memory;
- to be included in cues to episodic memories;
- to come to consciousness;
- to be written to episodic memory as parts of events, and
- to be available for the creation of new schemes by the procedural learning mechanism.

This kind of implementation would give such actions first-class status among the ontological entities of the LIDA model. Self-as-agent would then be realized as the set of all self-action nodes in the slipnet.

Expectations codelets are a specific type of attention codelets that are produced with every action selected in LIDA. The expectation codelet attempts to bring to consciousness items in the workspace that bear on the success of the given action achieving its expected result. Thus LIDA's expectation codelets will be part of the self-as-agent implementation.

Self-as-subject can be implemented as the set of acted-upon nodes in the slipnet, i.e., nodes representing actions by other entities upon the agent such as being pushed, stroked, hugged, slapped, yelled-at, fallen-upon, etc.

Self-as-experiencer might be thought of as being comprised of all of the rest of the slipnet. The Minimal Self can be implemented simply from the existing modules in the LIDA model.

Implementing Extended Self: Here we consider the four parts of the Extended Self from Figure 1. The Autobiographical Self is the collection of episodic memories of events that one has about himself or herself, rather than only about others. These memories have to have come from consciousness. In LIDA, the local associations from transient episodic memory and declarative memory come to the workspace in every cognitive cycle. This requires a verifiable report (of that memory coming to consciousness). Not all of them may be operationally verifiable.

The Selfplex is personal beliefs and intentions. In the LIDA model, the agent's beliefs are in the semantic memory. Intentions are represented by the intentions codelets. These are processes that get generated at each volitional goal selection. They look for opportunity to bring information concerning the goal to the Global Workspace. In LIDA, each volitional goal has an intention codelet.

Action that is taken volitionally, that is, as the result of conscious deliberation, is an instance of the action by the Volitional Self. Deliberate actions occur in LIDA and are represented as behavior streams. Thus LIDA has a volitional self. Deliberative acts have to be conscious, in the sense that the process of deliberation has to be conscious before the act itself.

An action to be influenced by the Narrative Self must intend to convey something meaningful about the speaker; it can be

determined by the presence of either explicit or implied personal pronouns. First, a LIDA-based agent has to understand such self-report requests. This can be implemented in the perceptual associative memory using perception codelets, slipnet and working memory. Then the agent has to generate the reports based on its understanding of such requests. The LIDA model facilitates this with existing modules. A LIDA-based agent can have motivations to report on itself and enjoy responding to such queries about itself, with feeling nodes in its perceptual associative memory. The agent has to become conscious of such a request, by its attention codelets, specifically built for such a task. We need reporting behavior streams in the procedural memory that can generate reports from the contents of consciousness.

Effectively, the LIDA model provides for the basic blocks to implement the various parts of a multi-layered self system as hypothesized in GWT. There are several interesting issues that such an implementation would bring up, which we will look at in the discussion section of this paper.

5 DISCUSSION

The main goal of our research work is to understand how the mind works. Implementing a self system in the LIDA model provides a better and more complete understanding of cognition and the Global Workspace Theory.

We see that the Proto-Self is already part of the LIDA model and is not built as a separate module/structure. This may be the case with most cognitive software agents/cognitive robots. The very nature of these systems requires the global parameters for the functioning of these agents, thus affecting the state of the software agent or robot.

In contrast, the Minimal/Core Self and the Extended Self need to be implemented in the LIDA model. While the Minimal Self can be easily facilitated in the LIDA model with the existing modules, the Extended Self requires new structures to be added to the existing modules. Implementing the various pieces of the self system would take us one step closer to a complete model of cognition.

An autonomous agent/cognitive robot based on the LIDA model that also has a self system might be suspected of being close to subjectively conscious for several reasons. First, such an agent/robot would be functionally conscious. Further, it could be made to fulfil the coherent, stable perceptual world condition [14]. We claim that such an agent/robot will take us one step closer to realizing phenomenal consciousness in these cognitive models.

Today researchers at the Brain Mind Institute at EPFL are using virtual reality and brain imaging to understand how the human body is represented in the brain and how this affects the conscious mind [37]. The self system is directly linked to consciousness and as we implement models of machine consciousness, it is imperative that we include the self system in these models.

REFERENCES

[1] Baars, Bernard J. (1988), *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.

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

[3] Baars, B J. (2003), How brain reveals mind: Neural studies support the fundamental role of conscious experience. *Journal of Consciousness Studies* 10: 100–114.

[4] Baars, Bernard J and Stan Franklin (2009). Consciousness is Computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness*, 1(1) : 23-32.

[5] Baars, Bernard J and Stan Franklin (2003). How conscious experience and working memory interact. *Trends in Cognitive Science* 7: 166–172.

[6] Baars, B J, T Ramsoy, and S Laureys (2003). Brain, conscious experience and the observing self. *Trends Neurosci.* 26: 671–675.

[7] Baddeley, Alan, Martin Conway, and John Aggleton (2001), *Episodic memory*. Oxford: Oxford University Press.

[8] Blackmore, Susan (1999). *The meme machine*. Oxford: Oxford University Press.

[9] Damasio, Antonio R (1999). *The feeling of what happens*. New York: Harcourt Brace.

[10] Franklin, S, B J Baars, U Ramamurthy, and Matthew Ventura (2005). The role of consciousness in memory. *Brains, Minds and Media* 1: 1–38.

[11] Franklin, S. and Ramamurthy U. (2006), Motivations, Values and Emotions: 3 sides of the same coin. *Proc. 6th International Workshop on Epigenetic Robotics*, Paris, September 2006, Lund University Cognitive Studies, 128: 41-48.

[12] Gallagher, Shaun (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Science* 4: 14–21.

[13] Gazzaniga, Michael S (1998). *The mind's past*. Berkeley: University of California Press.

[14] Merker, Bjorn (2005). The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. *Consciousness and Cognition* 14: 89–114.

[15] U Ramamurthy, B J Baars, S K D'Mello and S Franklin (2006), LIDA: A Working Model of Cognition. *Proc. 7th International Conference on Cognitive Modeling*, 244-249.

[16] Seth, A K, B J Baars, and D B Edelman (2005). Criteria for consciousness in humans and other mammals. *Consciousness and Cognition* 14: 119–139, (2005).

[17] Shanahan, M P (2006). A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and Cognition* 15: 433-449.

[18] Strawson, G. (1999). The self and the sesmet. In *Models of the self*, ed. Shaun Gallagher and J Shear: 483–518. Charlottesville, VA: Imprint Academic.

[19] Goodale, M. A., and D. Milner (2004). *Sight Unseen*. Oxford: Oxford University Press.

[20] Varela, F. J, Thompson, E., & Rosch, Eleanor (1991). *The embodied mind*. Cambridge, MA: MIT Press.

[21] Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577–609.

[22] Conway, M. A. (2001), Sensory-perceptual episodic memory and its context: Autobiographical memory. In A. Baddeley, M. Conway, & J. Aggleton (Eds.), *Episodic memory*. Oxford: Oxford University Press.
[23] Baddeley AD, Hitch GJ. (1974), Working memory. In Bower GA (Ed), The Psychology of Learning and Motivation. New York: Academic Press, pp 47–89.
[24] Glenberg A. M. (1997) What memory is for. *Behavioral and Brain Sciences* 20:1–19.
[25] Ericsson KA, Kintsch W. (1995). Long-term working memory. *Psychological Review* 102: 211–245.
[26] Sloman A. (1999), What Sort of Architecture is Required for a Human-like Agent? In Wooldridge M, Rao AS (eds), *Foundations of Rational Agency*. Dordrecht: Kluwer Academic Publishers, pp 35–52.
[27] Hofstadter DR, Mitchell M. (1995), The Copycat Project: A model of mental fluidity and analogy-making. In Holyoak KJ, Barnden JA (eds), *Advances in connectionist and neural computation theory, Vol. 2: logical connections*. Norwood N.J.: Ablex, pp 205–267.
[28] Kanerva P (1988) *Sparse Distributed Memory*. Cambridge MA: The MIT Press.
[29] Maes, P. (1989), How to do the right thing. Connection Science 1: 291–323.
[30] Marshall, J. (2002), Metacat: A self-watching cognitive architecture for analogy-making. In *24th Annual Conference of the Cognitive Science Society*:631-636.
[31] Drescher, Gary L. (1991). *Made-up minds: A constructivist approach to artificial intelligence.* Cambridge, MA: MIT Press, (1991).
[32] Brooks RA. (1991), How to build complete creatures rather than isolated cognitive simulators. In VanLehn K (ed), *Architectures for Intelligence*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp 225–239.
[33] Chaput, Harold H., Benjamin Kuipers, and Risto Miikkulainen (2003). Constructivist learning: A neural implementation of the schema mechanism. In *Proceedings of WSOM '03: Workshop for Self-Organizing Maps.* Kitakyushu, Japan.
[34] Stan Franklin, Uma Ramamurthy, Sidney K. D'Mello, Lee McCauley, Aregahegn Negatu, Rodrigo Silva L., and Vivek Datla (2007). LIDA: A Computational Model of Global Workspace Theory and Developmental Learning, *AAAI 2007 Fall Symposium - AI and Consciousness: Theoretical Foundations and Current Approaches*, Menlo Pk, CA: AAAI.
[35] Baars, Bernard J and Stan Franklin, (2009). Consciousness is computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness*, (2009).
[36] Uma Ramamurthy, Bernard J. Baars, Sidney K. D'Mello, and Stan Franklin (2006). LIDA: A Working Model of Cognition. *7th International Conference on Cognitive Modeling*, Eds: Danilo Fum, Fabio Del Missier and Andrea Stocco, p. 244-249.
[37] The Science of Self – Philosophy and Neurobiology: http://actualites.epfl.ch/newspaper-article?np_id=1648&np_eid=114
[38] The real avatar: Researchers use virtual reality and brain imaging to hunt for the science of the self: http://www.physorg.com/news/2011-02-real-avatar-virtual-reality-brain.html

Consciousness, Meaning and the Future Phenomenology

Ricardo Sanz[1,2], Carlos Hernández[1] and Guadalupe Sánchez[1]

Abstract. Phenomenological states are generally considered sources of intrinsic motivation for autonomous biological agents. In this paper we will address the issue of exploiting these states for robust goal-directed systems. We will provide an analysis of consciousness in terms of a precise definition of how an agent "understands" the informational flows entering the agent. This model of consciousness and understanding is based in the analysis and evaluation of phenomenological states along potential trajectories in the phase space of the agents. This implies that a possible strategy to follow in order to build autonomous but useful systems is to embed them with the particular, ad-hoc phenomenology that captures the requirements that define the system usefulness from a requirements-strict engineering viewpoint.

1 INTRODUCTION

Research into machine consciousness is justified in terms of the potential increase of functionality [25] but also as a source of experimentation with models of human consciousness to evaluate their value [19].

Even when there are old arguments against the possibility of machine consciousness[3], several attempts at realisations of machine consciousness have been done recently [19]. In some cases, these systems propose a concrete theory of consciousness explicitly addressing artificial agents [15, 10] but in other cases the implementations follow psychological or neural theories of human consciousness developed without considering machines as potential targets for them. This is true, for example in the case of the many implementations of Baars' Global Workspace Theory of consciousness [3, 21, 13, 26].

These are very valuable efforts that help clarify the many issues surrounding consciousness and foster a movement towards making more precise the sometimes too-philosophical terms used in this domain. All these different implementations —if accepted as conscious— may be considered as exemplars in an attempt towards an ostensive definition of *consciousness* that includes humans and maybe also some animals [4].

However as pointed out by Sloman [28] *"pointing at several examples may help to eliminate some misunderstandings by ruling out concepts that apply only to a subset of the examples, but still does not identify a concept uniquely since any set of objects will have more than one thing in common."* In a sense, the only possibility of real, sound advance in machine consciousness is to propose and risk a background theory against to which experiments are done and evidence thrown. This is indeed the path followed by the works previously mentioned of Chella, Haikonen, Franklin, Arrabales or Shanahan. However, most of the approaches are focused on just one aspect of consciousness [5]. The multifarious character of consciousness is an obvious problem.

Indeed, Sloman [28] suggests that the main difficulty that we confront in the research on consciousness and machine consciousness is related to the *polymorphic* nature of the *consciousness* concept. This may seem to imply that trying to tackle several aspects of consciousness -access consciousness, phenomenal consciousness, self-awareness, etc. — in one single shot —a single model, a single robot— is hopeless. This program of addressing consciousness as a whole is also hampered by the semantical flaws that some of the conceptions of consciousness suffer when abstracted from specific contexts.

However, Sloman also recognises that *"perhaps one day, after the richness of the phenomena has been adequately documented, it will prove possible to model the totality in a single working system with multiple interacting components."* This is, boldly, what we try to do inside our long term ASys research program. In order to progress in the systematic engineering of autonomous, robust agents, we will try to make them conscious. And will try to do so by using a *single*, *general* and *unified* theory of consciousness[4].

The approach taken in this effort directly attacks the polymorphic nature of the concept. We will express general consciousness mechanisms in the form of architectural patterns that will be instantiated in the several forms that are necessary for the specific uses of a particular agent. This approach breaks up the unicity/variety problem of consciousness, leveraging a single structure for different uses.

2 THE REASONS FOR ACTING

The quest for control architectures for artificial autonomous agents confronts a problem concerning the relations between the goals of the agent and the goals of the owner. This is very much connected with the value systems of humans and how these drive their behaviour [23].

Phenomenological states are generally considered sources of intrinsic motivation for autonomous biological agents. At the end of the day, what counts is the phenomenology. What is relevant for the agent is how the internal changes concerning its perception of the world and of itself impacts its experiential state [9].

To be more precise, for us humans, what counts is the integral, i.e. an accumulated value, of the phenomenological states along the lived trajectories —past, present and future. This

[1] Autonomous Systems Laboratory, Universidad Politécnica de Madrid, José Gutierrez Abascal 2, 28006 Madrid, Spain. www.aslab.upm.es. Email: {Ricardo.Sanz, Carlos.Hernandez, Guadalupe.Sanchez}@aslab.org.

[2] Sackler Center for Consciousness Science, University of Sussex, Falmer, East Sussex, UK. www.sussex.ac.uk/sackler.

[3] Paul Ziff, in 1959 said: "*Ex hypothesi* robots are mechanisms, not organisms, not living creatures. There could be a broken-down robot but not a dead one. Only living creatures can literally have feelings." [32]

[4] *Single*, because we are going to propose only one; *general* because we intend it to be of applicability to any kind of system, whether natural or artificial; and *unified* because it shall address all the conceptual spectrum of consciousness (except bogus terms).

is the very foundation for acting —the reasons to act— and the very grounding of ethics. We just care about feeling well and having the right experiences. This may sound a bit selfish but even altruistic behaviour shall be gratifying in some sense (albeit, if this is right, in a phenomenological sense).

This position will be clarified later in terms of what it means saying that the phenomena are the source of all behaviour. To do this we must enter into an analysis of the nature of meaning and consciousness. Both in natural and artificial settings.

Following a general approach is necessary for the objective of the ASys program of targeting a universal theory of consciousness —in terms of enabling the construction of better autonomous systems— but it is also of maximal relevance when addressing the construction of systems interacting with humans. In order to provide machines suitable for interacting with humans' lives —and most machines are designed to do so— it is necessary to understand this phenomenological grounding for action in humans and also it may be necessary to investigate the possibilities of such a phenomenological stance concerning the realization of machines.

3 ABSTRACT ARCHITECTURE OF A CONSCIOUS MACHINE

Our strategy in the search for a general architecture for consciousness is based in the identification of a set of architectural principles that will guide the definition of reusable design patterns [7]. An early version of these principles was presented in [25]. These principles offer precise but general definitions of some critical concepts in mind theory (like *representation*, *perception*, *action*, *value*, *consciousness*, etc.).

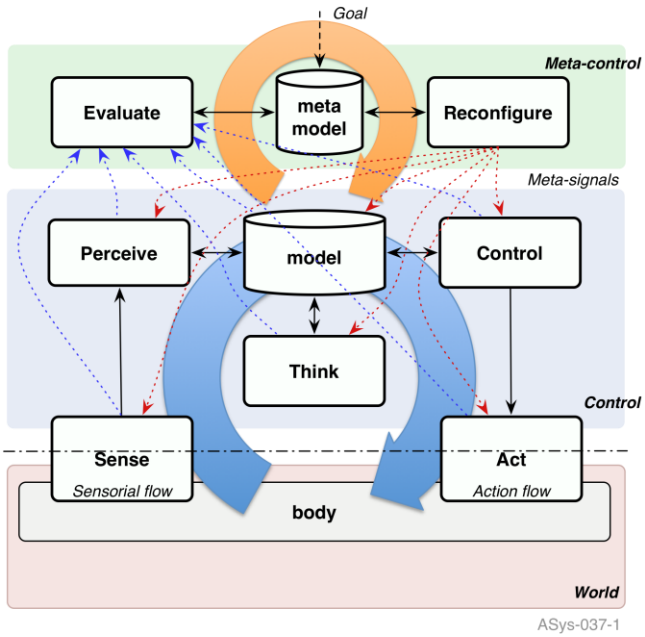

Figure 1. The basic building blocks for a design and realisation of a conscious machine are polymorphic patterns. The figure shows two of the basic patterns used in the definition of the cognitive architecture of reference for general consciousness: EPISTEMICCONTROLLOOP and METACONTROL.

The current set of design principles is the following:

1. *A cognitive system builds and exploits models of other systems in their interaction with them*. These models are —obviously— representations. They sustain the realisation of a model-based control architecture. Models are made at multiple levels of resolution and may be aggregated to constitute integrated representations.
2. *An embodied, situated, cognitive system is as good performer as its models are*. The ideal condition is achieving isomorphism in a certain modelling space. It is important to note that models are always abstractions hence defining a modelling space that is inherently different from that of the modelled system.
3. *Except in degenerate cases, maximal timely performance is achieved using predictive models*. What counts for an agent is the value got not only now, but from now on up to a fuzzy time horizon. The depth of the horizon will be dependent of the specific aspect that is anticipated.
4. *Perception is the continuous update of the integrated models used by the agent by means of real-time sensorial information*. Perceiving is hence much more than sensing. Sensing is the mapping of physical estates of the sensed entity into informational states inside the perceiving agent. In a second stage perceptual mechanics updates/creates models to exploit this information. Note that models are necessarily based on a sustaining ontology. This implies that perception suffers model-related ontological blindness.
5. *Agents perceive and act on the basis of multiple integrated, scalable, unified models of task, environment and self*. Model-based control is the core mechanism for action generation. This enables a search for global performance maximisation (obviously bounded by what is known/modelled). Model and action integration may happen at multiple scales.
6. *An aware system is continuously perceiving and computing meaning from the continuously updated models*. Meaning is defined as the partitioning of state-space trajectories in terms of value for the agent. What is different in this proposal for a concept of meaning is that we are considering not only the current state of affairs but the potential future values for the agent.
7. *Models are executed by engines and may be collapsed with them into simpler subsystems*. Model execution leverages models in the obtainment of many classes of data of relevance to the agent: actions, states, causes, means, etc. Model execution is hence necessarily continuous, multiple —forward, backward, means-ends, etc.— and concurrent. In some cases models and engines may be collapsed into a simple, more efficient element. Model-engine collapses are efficiency-exploitability tradeoffs. Collapsed models sacrifice multiple use to gain effectiveness.
8. *Attentional mechanisms allocate both physical and cognitive resources for system perceptive and modelling processes so as to maximize performance*. The bandwidth of the sensory system is enormous and the perceptual task is not easy. The amount of sensed information that may be integrated in the mental models of the agent is bounded by the availability of resources. The allocation of resources to subsets of sensed information is done using cognitive control and also immediate anticipatory valuation (significance feedback).

Note that this implies a primary form of perception before the conscious level.

9. *The agent reconfigures its functional organisation for context-pertinent behaviour using value-driven anticipatory metasignals*. This is the role played by (some) emotional mechanisms [24].
10. *A self-aware system is continuously generating meanings from continuously updated self-models*. The agent perceives and controls itself as it perceives and controls the world. "Self" is the closure of the executing self-model.

These principles are being reified in the form of design patterns (see Figure 1) and implemented using state of the art object-oriented software technologies.

This pattern-based approach enables the formerly stated vision of having both a general approach and the concrete implementations necessary for the diversity of tasks that an agent must address.

In this line of work, Hernández has proposed The Operative Mind (OM) [17] as an architectural framework for development of bespoke systems. This class of architectural reference model —in the line of RCS [1] or CogAff [29]— can be used for engineering systems which implement, as we claim, analogue functional capabilities to those reported —top-down causality, flexible control, integration, informational access, and intrinsic motivation— of biological consciousness. This enables, as a result, improved autonomy and robustness.

Figure 2. The Higgs robot is the experimental platform used for the deployment of the OM Cognitive Architecture.

Consciousness is implemented on it as a set of services, in an operating system fashion, based on deep modelling of its own control architecture [18], that supervises the adequacy of its structure to the current objectives in the given environment [20] triggering and managing adaptivity mechanisms. This system is being implemented in the control system of an autonomous mobile robot (see Figure 2).

4 MODEL-BASED PREDICTIVE CONTROL AND PHENOMENOLOGY

The architectural model proposed in the above principles is consonant with the model-based control strategies used in technical environments —industrial plants, aircraft, etc. [8].

In model-based predictive control (MBPC), the controller produces the next instantaneous action by i) first projecting a desired trajectory of targets optimised for that goal, ii) then predicting the future consequences of the actions needed to follow that trajectory to obtain precisely an optimised plan of actions, and finally iii) executing only the first action in the plan; then the cycle starts over again.

Notice that for step i) a cost function is used, which is both a model of the task and an evaluation procedure, and for ii) a model of the plant –i.e. system (body) and environment– is employed.

So far, control systems based on advanced techniques such as MBPC contain informational structures and processes that our model could ascribe to access consciousness: they exploit updated models of the plant and evaluate in the view of the predicted future. But as far as the model do not include the system itself –i.e. the controller–, the system is not self-conscious. This implies that there are no phenomenological states concerning the own agent involved.

Now let us suppose that the system/controller includes a model of itself, so it evaluates not only the future environment states given its possible actions, but also its very own possible future states. Then we will have a system that, from sensory information flow, would generate informational structures containing an evaluation of its processing, not only current, but as predicted in the future according to its past.

It is important to note that the evaluation is realised in terms of the value obtained by the agent. In the case of artificial control systems these values are imposed by externally grounded utility functions. In the case of biological systems these utility functions are internal and expressed in terms of what is good and bad for the agent: i.e. its experience. The metaperception of the agent as perceiver sustains the valuation of goodness of states. This may constitute the very substrate of phenomenology: the system, by virtue of the described process, would be *experiencing* that sensory input.

The grounding of experience on model-based metaperception provides an operational understanding of the "what is it like to be" question [22]. To know what is it like to be a bat would require not only the echolocation sensory system but the full perceptual pipeline and the metaperceptual pipeline. We cannot experience being a bat if we don't meet these requirements, but, however, we can have a deep theory of what it is like to be a bat and hence know "what is it like to be it".

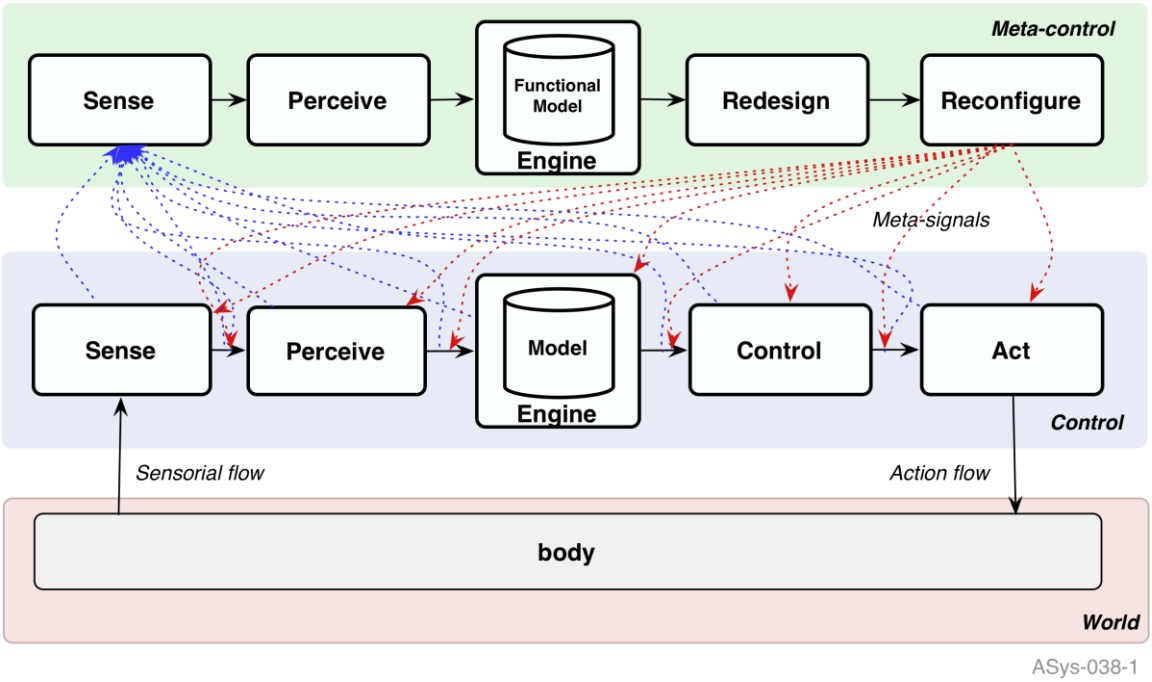

Figure 3. The self-perception, self-configuration meta-loop shares the patterned structure of the EPISTEMICCONTROLLOOP. The meta-level gathers information about the functional organisation of the lower epistemic control loop and may act to change it. The observed/controlled world of the metaloop is a functioning cognitive agent.

Note that the action part of the meta loop shown in Figure 3 shows action modifying the workings of the lower, world-situated loop. The meta-control competences enabled by self perception constitute the active part of emotional mechanisms [24]. In a sense, consciousness, meaning and emotion are stepping-stones in the same road [2].

5 MEANING AND THE FUTURE

In this paper we provide an analysis of 'consciousness" in terms of a precise definition of how an agent "understands" the informational flows entering the agent. This definition of understanding is based in the analysis and evaluation of phenomenological states along potential trajectories in the phase space of the agents.

We propose a rigorous definition of "meaning" in terms of the separation of potential agent trajectories in different value classes —consider that the information flows are a critical resource for trajectory enaction and separation. The values to be computed will not be in the particular space of magnitudes of an external, third person observer but in the magnitudes of relevance to the agent: i.e. the phenomenological ones. This computation requires from the agent an intrinsic capacity for anticipation —including anticipation of phenomenological states.

Note that in this context *phenomenological* is not restricted to the limited interpretation in terms of qualia, but in the broader sense of *phenomenal structure* [30]:

"the phenomenal structure of experience is richly intentional and involves not only sensory ideas and qualities but complex representations [our models] of time, space, cause, body, self, world and the organized structure of the lived reality"

For the reasons stated before, this model —of meaning and consciousness— shall be of applicability both to humans and robots, hence implying a rigorous analysis and definition of phenomenological states —because rigour is necessary if this is going to be built into the robots and not just predicated from some externally observed behaviour.

Clarifying these issues is not only of relevance for robot construction but also for advancing into a general theory of consciousness both operational in the technological side and explanatory in the biological one —e.g. being useful to create safer machines [25] and being able to explain the nature of pain asymbolia [14].

Consider the situation of a system at certain time (now, t_0) where the system must decide what to do based on a certain information it has received (see Figure 4). The system has followed a certain trajectory $x(t)$ in its state space but the future is open concerning the different possibilities for acting (A_a, A_b, A_c). The concrete future trajectory will depend on the concrete action, but will also depend on the concrete state of the world and the agent at t_0. The meaning of a piece of information —about the world or about the agent itself– is the way it partitions the set of possible future trajectories in terms of anticipated phenomenological states.

How is this meaning enacted? By integration of the information received into the model that the agent uses to predict

the future and by executing this model in forward time. In a sense, grasping the meaning of some information is leveraging this information in enhancing the prediction of how reality is going to behave.

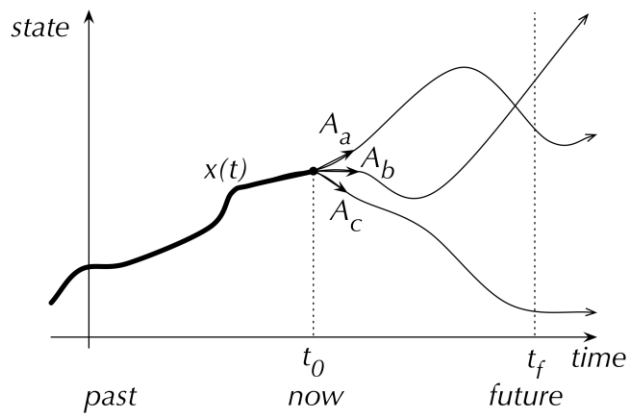

Figure 4. Understanding sensory flows and the derived emotional processes are strongly related to the anticipatory capabilities of the agents.

This interpretation of meaning and consciousness is indeed not new. As Woodbridge said [31] in relation to potential definitions of consciousness [6]: *Professor Bode states the general problem tersely, it seems to me, when he asks, "When an object becomes known, what is present that was not present the moment before ?" I have attempted to answer that question in one word — "meaning."*

Phenomenology goes beyond the experiential qualities of sensed information. Haikonen argues that *qualia are the primary way in which sensory information manifests itself in mind* [16] but in our model this qualitative manifestation is not necessarily primary but may be produced in downstream stages of the perceptual pipeline. What is important for us is not just the qualities of the sensed but the experience of their meaning. As Sloman and Chrisley [29] say, "an experience is constituted partly by the collection of implicitly understood possibilities for change inherent in that experience."

It must be noted that the model proposed is concurrent. This implies that the perceptual pipeline is operating in several percepts at the same time. But due to the integrated nature of the models —principle 5— these pipelines may eventually converge (in non pathological cases). This may imply a reduction of the focus of inner attention to a single percept. This is in line with Dennett's multiple drafts theory of consciousness [11].

6 CONCLUSSIONS: IS HETEROPHENOMENOLOGY A NEED ?

Going back to the analysis done at the very beginning of the paper on the construction of autonomous systems, and after describing the architectural picture of the ASys model of autonomy and consciousness, we reach the conclusion that heterophenomenology is a need.

However, heterophenomenology (phenomenology of others different from oneself) must be understood in a sense a bit different from the initial proposal of the term by Dennett [12] of using verbal reports (and other types of acts) as objective, third-person observations that provide the observer with partial information about the agent's beliefs regarding its own conscious experience.

In this context, building machines that experience, the problem of engineering the right phenomenological mechanism is crucial because it will be the origin of the intrinsic motivations of the agents. We must adopt an heterophenomenological engineering approach in the sense of being able to engineer phenomenologies into machines to match our very own needs [33]. These will not be human phenomenologies but the phenomenologies that when deployed will make the agents pursue our satisfaction.

But for this, we need not only a better understanding of the artificial [27] but of our own consciousness.

7 ACKNOWLEDGEMENTS

We acknowledge the support of the European Commission through Grant *HUMANOBS: Humanoids that learn Socio-communicative Skills by Imitation.*

REFERENCES

[1] James S. Albus (1991), 'Outline of a theory of intelligence', *IEEE Transactions on Systems, Man and Cybernetics*, 21(3), 473–509.

[2] Yuri I. Alexandrov and Mikko E. Sams (2005), 'Emotion and consciousness: ends of a continuum', *Cognitive Brain Research*, 25, 387 – 405.

[3] Bernard J. Baars (1997), 'In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness.', *Journal of Consciousness Studies*, 4, 292–309.

[4] Xabier Barandiarán and Kepa Ruiz-Mirazo (2008), 'Modelling autonomy: Simulating the essence of life and cognition', *Biosystems*, 91(2), 295–304.

[5] Ned Block (1995), 'On a confusion about the function of consciousness', *Behavioral and Brain Sciences*, 18, 227–247.

[6] B. H. Bode (1908), 'Some recent definitions of consciousness', *Psychological Review*, 15, 255–264.

[7] Frank Buschmann, Regine Meunier, Hans Rohnert, Peter Sommerlad, and Michael Stal (1996), *Pattern Oriented Software Architecture. A System of Patterns*, John Wiley & Sons, Chichester, UK.

[8] Eduardo F. Camacho and Carlos Bordons (2007), *Model Predictive Control*, Springer, second edn.

[9] Peter Carruthers (2000), *Phenomenal Consciousness*, Cambridge University Press.

[10] Antonio Chella, Marcello Frixione, and Salvatore Gaglio (2008), 'A cognitive architecture for robot self-consciousness', *Artificial Intelligence* in Medicine, 44, 147–154.

[11] Daniel Dennett (1991), *Consciousness Explained*, Penguin.

[12] D.C. Dennett (2003), 'Who's on first? heterophenomenology explained', *Journal of Consciousness Studies*, 10, 19–30.

[13] Stanley P. Franklin (2000), 'Building Life-Like 'Conscious' Software Agents', *Artificial Intelligence Communications*, 13, 183–193.

[14] Nicola Grahek (2007), *Feeling Pain and Being in Pain*, MIT Press, second edn.

[15] Pentti O. Haikonen (2003), *The Cognitive Approach to Conscious Machines*, Imprint Academic, Exeter.

[16] Pentti O. Haikonen (2009), 'Qualia and conscious machines', *International Journal of Machine Consciousness*, 1(2), 225–234.

[17] Carlos Hernández, Ignacio López, and Ricardo Sanz (2009), 'The operative mind: a functional, computational and modelling approach to machine consciousness', *International Journal of Machine Consciousness*, 1(1), 83–98.
[18] Owen Holland and Ron Goodman (2003), 'Robots with internal models - a route to machine consciousness?', *Journal of Consciousness Studies*, 10(4-5), 77–109.
[19] Lyle N. Long and Troy D. Kelley (2009), 'The requirements and possibilities of creating conscious systems', in Proceedings of the *AIAA InfoTech@Aerospace Conference*, Seattle, USA.
[20] Ignacio López (2007), *A Framework for Perception in Autonomous Systems*, Ph.D. dissertation, Departamento de Automática, Universidad Politécnica de Madrid.
[21] Raúl Arrabales and Araceli Sanchis, (2008), 'Applying machine consciousness models in autonomous situated agents', *Pattern Recognition Letters*, 29(8), 1033–1038.
[22] Thomas Nagel (1974), 'What is it like to be a bat?', *The Philosophical Review*.
[23] Michael Pauen (2006), 'Emotion, decision, and mental models', in *Mental Models and the Mind*, eds., Carsten Held, Markus Knauflf, and Gottfried Vosgerau, Elsevier.
[24] Ricardo Sanz, Carlos Hernández, Jaime Gómez, and Adolfo Hernando (2010), 'A functional approach to emotion in autonomous systems', in *Brain Inspired Cognitive Systems 2008*, eds., Amir Hussain, Igor Aleksander, Leslie S. Smith, Allan Kardec Barros, Ron Chrisley, and Vassilis Cutsuridis, volume 657 of Advances in Experimental Medicine and Biology, 249–265, Springer, New York.
[25] Ricardo Sanz, Ignacio López, and Julita Bermejo-Alonso (2007), 'A rationale and vision for machine consciousness in complex controllers', in *Artificial Consciousness*, eds., Antonio Chella and Riccardo Manzotti, Imprint Academic.
[26] Murray Shanahan (2006), 'A cognitive architecture that combines internal simulation with a global workspace', *Consciousness and Cognition*, 15(2), 433–449, (June 2006).
[27] Herbert A. Simon (1996), *The Sciences of the Artificial*, MIT Press, Cambridge,USA, 3rd edition.
[28] Aaron Sloman (2010), 'Phenomenal and access consciousness and the "hard" problem: A view from the designer stance', *International Journal of Machine Consciousness*, 2(1), 117–169.
[29] Aaron Sloman and Ron Chrisley (2003), 'Virtual machines and consciousness', *Journal of Consciousness Studies*, 10(4-5), 133–172.
[30] Robert van Gulick (2004), 'Consciousness', in *Stanford Encyclopedia of Philosophy*, Stanford University.
[31] Frederick J. E. Woodbridge (2006), 'Conciousness and meaning', *Psychological Review*, 15(6), 397 – 398, (1908). University of Aberdeen, UK.
[32] Paul Ziff (1959). The feelings of robots. *Analysis* 19(3), January 1959: 64-68.
[33] Ron Chrisley (2009), Synthetic phenomenology. *International Journal of Machine Consciousness*, 1:53–65.

Can Functional and Phenomenal Consciousness Be Divided?

John G Taylor
Department of Mathematics, King's College, Strand,
London WC2R 2LS UK.
Email: taymore2002@aol.co.uk; **Web:** http://www.raceforconsciousness.co.uk

Abstract. We answer the question raised by the title by developing a neural architecture for the attention control system in animals in a hierarchical manner, following what we conjecture is an evolutionary path. The resulting evolutionary model (based on CODAM at the highest level) and answering the question allows us both to consider different forms of consciousness as well as how machine consciousness could itself possess a variety of forms.

Keywords: Attention; Evolution; Levels of Attention; Levels of Consciousness; Machine Consciousness

1 INTRODUCTION

Darwin's magisterial 'Descent of Man' [1], supported by many writers on evolution since, implies that consciousness has decided functionality in humans. Yet some have proposed that consciousness is purely an epiphenomenon, with no real purpose, but arising as a necessary evil from the depths of decision making components of the brain. It could function as a checking device, as suggested by Benjamin Libet [2] from his well-known experiments on when consciousness arises in subjects as they are about to act. His result that consciousness only arose after the beginning of the developing action (as seen by the 'bereitspotential') seemed to underline the lack of value of the ensuing consciousness. It even led to suggestions to train children and artists to reduce the importance of consciousness in their lives and thereby become more creative. How can we reconcile these two views?

To begin to do that, we consider the concept of an inner self, as supposed to exist by many Western phenomenologists [3] and earlier by Kant, even, it has been suggested going back to Aristotle as 'a bit on the side'. Without such an inner self or 'I' we are all zombies (as was claimed by Dennett as part of his attack on qualia [4]). But does the inner self really exist? To begin to answer this difficult question we will trace a possible development of attention across the millions of years of man's ancestors, until we arrive at ourselves.

The question of the existence of an inner self may also be suggested as a crucial one for machine consciousness research. If human consciousness has such an important component then how can it be claimed that a similar component is not also crucial for a supposedly conscious machine? Indeed, from the approach of the evolution of consciousness, it can be seen that without such a component of 'inner self' such a machine would at best be a zombie. It would have no experience of any internal 'mental' activity. In the usual sense of the word, in other words it would not have a mind. To use such a finer sense of the word 'conscious', then, would not allow us to term a machine without such an inner self a 'conscious' machine. We conclude that such a 'bit on the side' is essential for a machine to be properly called 'conscious'. Hence there arises the relevance of this paper to a Symposium on Machine Consciousness. It should be added that following the evolutionary track to attention will allow us to see how a hierarchy of consciousness can be defined, although with considerable weakening of the meaning of the term 'conscious'.

In order to attack this question of the existence of an inner self through evolutionary development, we will explore how the process of attention could initially have been at a very primitive level of control, enabling lower animals to single out possible prey but without any necessary consciousness of the relevant stimuli. Such could be at the level of a crocodile, a mean predatory machine but one with little beyond its ability to launch a rapid attack on its prey.

We need to understand how this primitive attention control system could have been improved by addition of a working memory storage system, so allowing the resulting attended stimulus to be held for enough time to enable some form of primitive reasoning. At a further stage (or in parallel) of evolution, goal biasing could have been added to the primitive attention control. Finally the use of an attention copy signal would then allow more efficient processing by preventing distracters from getting in the way, as well as speeding up the overall process of attending to a given stimulus.

The resulting levels of sophistication of attention control are recognisable in lower animal species. Most crucially for ourselves, the highest and most sophisticated attention system has an important role for attention: that of speeding up attention to a target, and of reducing errors in attending to that target. A similar functionality of consciousness (as improving attention movement and reducing errors from distracters) could thereby become clear from this point of view, and leads to the non-triviality of consciousness as part of the most sophisticated attention control system of all species [5]. We need to explore this sophistication, and hence expose a possible mechanism for the creation of the inner self, in terms of the possibility of dividing consciousness into a functional and a phenomenal component.

In order to achieve such an interpretation of consciousness, we must consider how we can define attention so as to expose a more sophisticated level than that of the crocodile mentioned earlier or even that of the animals we have around us constantly, such as cats or dogs, or in the countryside the sheep, cows and horses which are the backbone of the farming economy.

That we begin to do in the next section, where a progression of control models of attention is introduced. It is on the basis of this sequence that we can conjecture as to a possible evolutionary progression of the detailed structure of attention, and most particularly how attention, regarded in enough detail, could contain in its interstices the glimmerings of consciousness which flower into full adult consciousness as an infant grows up. It is in these interstices that we must attempt to discover the more detailed structures which may exist in consciousness, and in particular if there is a natural separation of the overall brain architecture to allow us to recognise the possible two components of functional and phenomenological consciousness. At the same time we will lay the foundation for an approach to consciousness which also lays a basis for that to machine consciousness. For if we can describe consciousness in terms of a possible neural architecture then we can begin to explore how to implement it in a machine.

We start, therefore, in section 2, with a brief survey of attention as one of the simplest control systems, that of ballistic control, and a description of how such a simple control system can gradually be expanded to incorporate even more sophisticated neural architectures. In the following section 3 we analyse how this expanded structure can be expanded even further so as to be able to make attention even more efficient. It is this final architecture that we consider, in section 4, as that inside which it may be possible to observe the creation of consciousness, and especially of the inner self. In particular we conjecture that the separate components of phenomenal and functional forms of consciousness can exist, each playing their role in ensuring that consciousness is functionally as efficient as it can be in controlling the movement of attention and in allowing the attended stimulus representations to be attended to and processed to a higher level for cognitive actions to be taken. In section 5 we present evidence for such a view from brain imaging. We conclude in section 6 with a brief return to Darwin and his evolutionary ideas, and to the problem of consciousness in machine consciousness.

2 ATTENTION BEGINS

At the lowest level, attention is to be regarded as a filter on inputs to the brain, discarding those of no or insufficient interest and concentrating on those which are crucial to the survival of the animal containing the faculty of attention. At this lowest level, attention acts as a ballistic type of controller, in the manner of that done by a rifle in the hands of a sniper. They know their target, and will take aim so as to be able to send a bullet winging accurately at their target stimulus. In the case of attention at the ballistic level, a similar target stimulus is chosen by the attending animal, and its activation in the animal's brain is amplified to the detriment of activations arising from distracters around in the animal's environment. This process is clearly effective for an animal such as the crocodile and many others. It can attend closely to its prey and in so doing make its moves toward such prey more effective. There may be elements of the environment that get in the way, such as objects that need to be navigated round in order to reach its prey. In these cases the animal may be able to switch its attention momentarily to such objects, so as to move around them efficiently.

This clearly requires both an ability to possess control over attention in a so-called top-down manner, where the goal of the attention system has been set up from the start of any search process, as well as in a bottom-up manner. In the latter attention is able to be switched rapidly from one object to a new one which has unexpectedly arisen in the animal's environment. These two forms of attention (usually termed endogenous and exogenous, respectively) can be fused together if we consider how the bottom-up form can excite higher brain systems and temporarily act as a top-down goal [6].

In evolution there may have been first a development of a saliency map, to automatically pick out the most salient inputs in a scene. This would have been most relevant to exogenous attention control. The saliency map would have been very likely sited at the top of any visual processing hierarchy, and migrate upwards as the processing hierarchy evolved. In evolutionary time there would have been expansion of the processing hierarchy to develop ever more efficient processing of features (especially in vision). Ultimately the salience map could have migrated to the top of the processing region. It could then have become, with the evolution of a primitive form of working memory, a storage site for goals. There could thereby have evolved the ability to possess top-down goal-driven attention. At the same time exogenous attention, always of great survival importance, could have evolved to have fast access to the goal-driving sites (as observed in numerous studies [7 Fox et al], or more directly access and replace any top-down attention goal.

Having a ballistic form of attention allows an animal to exist more efficiently in a crowded environment. However it does not provide a very efficient mechanism of survival, since events occur over time, and if the memory of the attended stimulus disappears too quickly the attended stimulus activity in the animal's brain cannot help in the survival game. To get over this it is necessary to add a short-term memory (of a few seconds) to hold the attended stimulus activity in the animal's brain. Having such a short term memory (STM) clearly grants better survival chances to the animal as a result of being able to hold in memory a target either as prey or predator. Thus we reach at this point the extended ballistic attention model, as described for example in [8].

Such an extended ballistic control model has strong neuro-scientific support. The ballistic model proper has been observed by brain imaging in monkeys and humans [9; 10; 11]. The existence of an STM as a further component in the extended ballistic model has been observed also by brain imaging [12]. Its extension to the more complete working memory involving both frontal and parietal lobe activity (corresponding to goal holding in the former and STM in the latter) has been documented in many experiments.

We thus see that at least in present living animal species attention plays an important role. It has a set of levels of neural architecture that may be delineated as follows:

Level 1: Purely externally-driven attention, with a salience map leading to attention amplification of any target activity;

Level 2: Purely ballistic attention (both externally and internally driven), with only a short level of activity corresponding to the attended stimulus activity being present in the animal's brain. Such target-based activity could arise originally from a bottom-up target (as at level 1), or from a top-down (possibly long-term memory-based) target, so that this second level involved a competition for directing attention between the outside world and the inner brain environment;

Level 3: Extended ballistic attention, with an STM holding the activity;

Level 4: Doubly extended ballistic attention, with a goal being held over an extended period as well as the brain activity of the attended stimulus being held in an STM.

It is these four levels which we should be able to recognise in differing living animals, with the higher levels being present in higher-level animals. To also relate them to the evolutionary scheme of living things will be more difficult, due to the lack of soft tissue tracings in fossil brain remains. Thus, since only fossilised brain structures are available, only overall shapes of brains and their sizes will be available in the fossil domain. Living animals do provide some help in such a search, however.

3 IMPROVING ATTENTION

Attention is clearly a control system, from the evidence we have given above and from many other results presented both behaviourally and from brain imaging. Modern control theory has moved on considerably from ballistic control, however, as evinced by the many groups working in advanced control methods as applied to the control of industrial plants as well as to very complicated systems such as airplanes. In particular those which give a decided advantage to the system employing them might be expected, by Darwin's ideas of competitive evolution between species, to give such an advantage to a species employing them that such advanced control methods, arising from a mutation, would have been preserved in the competitive world.

Such a process has already been suggested as occurring in motor control by the brain [13]. Here the most important advance observed is that of the use by control systems of a predictor of the effects of the control action. This prediction is created by use of a copy of the motor control signal itself, by what is called a corollary discharge or efference copy of that signal. Such a predictor allows for two important features to be added to the control repertoire:

1) To correct for errors that might have arisen from the control signal so that these errors can be corrected early on, before a possibly incorrect target is reached, but after the control signal has been sent out (something clearly impossible in a ballistic control system);

2) Removal of distracters from possible attention, partly so as to avoid errors of target attention choice (as in 1)) as well as to prevent distracters from possibly accessing any STM or working memory sites so as to cause distraction in later use of attended brain activity.

The reason 2) involves us with the important question as to why attention? Having filtered out all but the attended stimulus content in the brain what is to be done with the resulting activity? It is expected to be coded at a very high level, but the use of an STM promotes efficiency if further processing is to occur. It is indeed such further processing, using the attended stimulus activity, that we consider (and know from our own experience) occurs in our brains. Activities such as thinking, imagining, reasoning and forming word sequences (or giving them meaning) are such activities. In particular since attention is heavily involved in these processes then we expect that it is the attended activity that is itself being employed in these high level activities.

The extension of attention by addition of a corollary discharge component was suggested in 2000 by the author [5] in the CODAM (for Corollary Discharge of Attention Model) CODAM was developed over the last decade in numerous ways, as in exploring the more detailed dynamics of attention control, of simulations of relevant experiments, and of exploring the nature of conscious experience as based on the CODAM approach [8; 14].

We present in figure 1 the overall architecture of CODAM, together with a specification of its components in the caption.

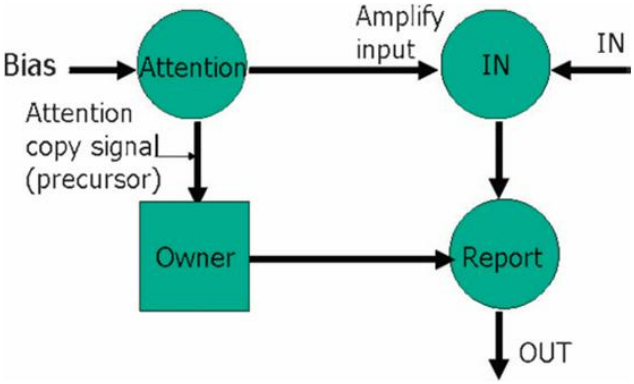

Figure 1. The Basic CODAM Architecture

Input enters the 'IN' module (possibly a hierarchy of neural modules), and attention is focussed on it to achieve its amplification, from some source of bias (bottom-up or top-down). The amplified activity is sent to a report module, acting as an STM for later processing.

The added module beyond the doubly-extended ballistic model is that of the STM for the corollary discharge of the attention movement control signal, termed in figure 1 the 'owner' module. The purpose of this further module is to provide the corollary discharge signal to be used for error correction and distracter removal on the input signal of the amplified representation of the attended target stimulus. This architecture is considered as the final stage in the evolution of consciousness as it arose in humans.

We thus reach a 5-stage developmental menu for the evolution of attention in animals. The fifth level corresponds to the addition of a corollary discharge of attention movement to the doubly-extended ballistic model at level 4. It is inside this hierarchy of attention control architectures, and especially

at the fifth level, that we must understand how consciousness could be created.

We will justify the use of our evolutionary approach so as to be able to separate out the phenomenological and functional components of consciousness. It is to that purpose we now turn.

4 CREATING THE INNER SELF

The CODAM model assumes that there exists some module, denoted 'Owner' in figure 1, which allows for a brief holding in short term memory of the corollary discharge signal. As assumed for report by means of the well-supported visual short-term memory (VSTM) acting as a receptacle for report of the content of an incoming target stimulus, so it is assumed that the content of the corollary discharge short term memory would also be available for similar report, although possibly for a briefer time. However the presence of the ownership signal of the corollary discharge of the attention movement signal gives this signal the only content, that of 'ownership' of the about-to-arrive visual stimulus into report of that content. There can be no other nature of the experience generated in the corollary discharge short-term memory, since the activity in that site is supposed not to be connected to lower level feature components enabling the stimulus activity to acquire content. Thus the owner activity is content free. But yet it possesses an experience of ownership due to the control it exerts over the access of the attended stimulus activity to its content report stage in the VSTM. Such control is assumed to consist of inhibition of possible distracters and amplification of the site for activation of the code for the attended stimulus.

The implication of this content-free but owned experience is that it can tentatively be identified with the 'inner self' of Western phenomenology of Husserl, Merleau-Ponty, Sartre and many other philosophers. Although these all had different detailed ideas on the inner or pre-reflective self they were all united as to its existence. The inner self has since been teased out more fully in detailed studies by [3] and by continued work on the process of loss of this content-free experience in schizophrenia [15; 8].

As originally proposed by Husserl [16] there is a specific timing sequence for the emergence of a conscious experience of content. This was supposed to be in three stages:

Pretention → Primal Impression → Protention

Each of these three stages was distinct: pretention arose at the early stage of the consciousness creation, the primal impression was that of the content of the attended stimuli, and protention involved a buffered memory of the experience.

We have earlier [17, 18] described how CODAM can explain these three temporal segments of the emergence of consciousness. Pretention is to be considered as the stages and associated experience involved in the creation of the attention feedback signal, the related corollary discharge activity and attention amplification of visual cortical activity representing the attended stimulus. The primal impression is the emergence of the amplified attended stimulus activity onto its buffer working memory for general report round the brain. Finally the protention period involves the continued but decaying activity on the buffer working memory site, the VSTM. Such a division of the dynamic activities in CODAM is a natural one, and fits nicely with the results of the experiential explorations of Husserl and his colleagues.

We can modify the temporal flow of experience from the above three component sequence so that the early processing under the heading of 'pretention' is now put under the different heading of 'ownership'. Such ownership involves the detailed control processes (inhibition and amplification) proposed for the corollary discharge signal and claimed above to have been observed in various paradigms [19; 20; 21].

In a manner similar to that in which the external world attains a constant form by means of the eye-movement corollary discharge [22], so we can expect that the ownership experience, that of the 'I', can be kept constant by means of its attention corollary discharge signal. This would thereby lead to what can be termed the 'Constant I', which is as directly experienced by each of us as we move through the world. The exact mechanism for this constancy is still unclear in the case of the external vision of the world (see [23] for a very recent discussion on this). In a similar manner we cannot conclude on a specific mechanism for the constant I. However we can expect there to be a close analogy between these two mechanisms from the analogy of the existence of the two corollary discharge mechanisms, the first for retinal movement and the second for attention movement.

In terms of the title of this paper, we can consider the period of attention involving the early dynamical activity, say from 180 – 400 msecs, as that involved in the phenomenological component of consciousness, whilst the further activity associated with holding of activity on the STM, with associated lower level features, as that of the functional component. Only the combination of the two will lead to consciousness itself, the earlier ownership activity being necessary to answer Nagel's question and give a sense of the experience of the activity gaining access to the STM in the functional phase. However from the CODAM viewpoint, it can be clear that all the above activity, in both the early and later stages, has important functions to play in the creation of consciousness itself.

5 EXPERIMENTAL EVIDENCE: The N2pc & the SPCN ERP SIGNALS

The corollary discharge component has been detected in ERP investigations of the dependence of the N2pc on masking: both forward and backward masking has been employed to study the effect on the N2pc [24]. The paradigm employed in [24] used a pair of coloured letters or digits, each presented for 100 msecs on either side of fixation (one digit and one letter were used at a time). The target character for detection had a specific colour, with one of the two characters presented having this colour, the other being the other colour (pink and green were the two colours employed). Immediately afterwards a second set of similar characters was presented for the same period, to act either as a backward mask or alternatively as a target, with the earlier pair of letters then functioning as a forward mask. In the no-mask case only one pair of letters was presented, with a blank screen for the second stimulus. The colour for a character category (letter or digit) was held constant for a given subject.

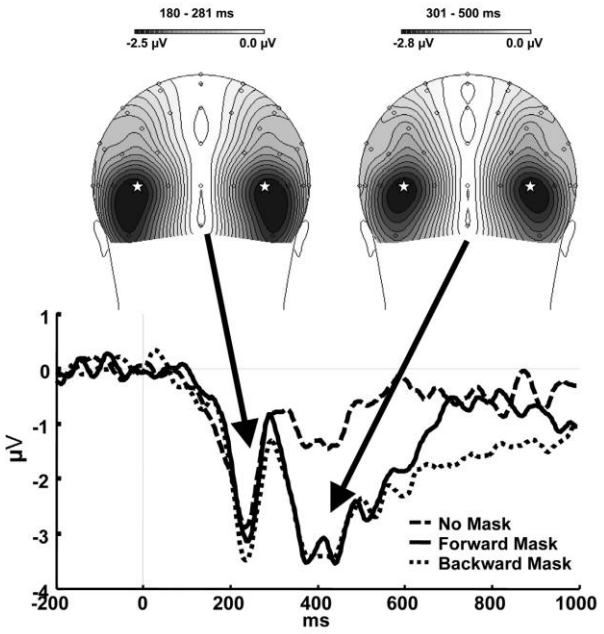

Figure 2. TheN2pc and SPCN as detected in [5].
The figure shows the N2pc (at 180-281 msecs) and the SPCN (at 301-900msecs) as observed in [24] (figure 3 of [24], permissions to be obtained).

The results of the experiment of [24] are given in figure 2, which shows there is an absence of any change of the N2pc caused by masking. Thus we can conclude that the N2pc (at least in the paradigm of [24]) is involved in focussing attention onto the relevant position in space for further processing to occur. In the masking paradigm used in [24] there was also the need to inhibit the distracter either coming just before (for forward masking) or just after (for backward masking) the target itself at the same position in space.

An important further component of processing, the SPCN, was observed [24] in the ERP signal over the period of 300 – 900msecs after stimulus onset, as shown in figure 6. The SPCN was longest for backward masking (from 300 – 900 msecs post-stimulus), shorter for forward masking (lasting for 300 – 700 msecs) and shortest for no masking (300 – 500 msecs). It was proposed by the authors that the SPCN reflected the presence of processing of the target and its mask (in the same hemisphere) inside the Short Term Memory (STM), this processing being shortest for the no-mask events. The dependence of the length of this processing indicated that the processing arose from removal of the distracter stimulus: this removal was more difficult for the backward masking case, less difficult for the forward mask and least difficult of all in the no mask situation.

The negativity aspect of the SPCN corresponds to removal of distracters, as does the N2pc at an earlier stage. Such inhibitory character of the N2pc has been proposed by numerous investigators [25; 26]. The distracters during the SPCN have penetrated the visual STM, as shown by continued activity of the SPCN, especially for backward masking. But then a corollary discharge of the attention signal, carrying attention goal information, must have been sent to the working memory module to provide suitable goal information in the STM. Such goal information is that of removing the distracter character activity in the VSTM in preference to that of the given target. The time during which the SPCN is acting appears appropriate for such processing in the working memory module (several hundred mille-seconds), with identification of this working memory site with the Short Term Memory site considered in [24].

We support this identification of the SPCN with a component of the corollary discharge by considering relevant details of the results presented in [24]: the SPCN carries a signature of the goal of the paradigm, as indicated by differences between the various temporal durations of the SPCN as correlated with the difficulty of the masking conditions and discussed in [24]. The SPCN involves goal-biased information in the parieto-occipital regions, as shown by SPCN activity being detected by MEG there [27]. Such information would arise from the intra-parietal sulcus/superior parietal lobe (IPS/SPL) source of the attention control signal [9] or directly from the goal module in prefrontal cortex, in prefrontal cortex/frontal eye fields (PFC/FEF) [9, 10, 11]. In either way we conclude:

The SPCN signal of [24] carries appropriate corollary discharge activity of attention movement.

It is possible to check the above by determining the correlation of the SPCN with the SPL/IPS and PFC/FEF activity. It would also be important to use Granger causality to show the causal flow of activity from the SPL/IPS or PFC/FEF sites so as to demonstrate that the SPCN, as a corollary discharge, is definitely arising in a causal manner from these latter sites; such data is not presently available.

The recent results obtained by using MEG [27], as noted, demonstrate that SPCN activity is sited in occipital and parietal regions of cortex, with the former expected to be involved in detail needed for the higher area activity in parietal. The latter will correspond both to the VSTM activity and to that of any maintenance sites. It is important to find any change in that distribution of activity, as seen by MEG, if the original masking paradigm in [24] were to be used. For such distracter activity is expected to be an important signature of any corollary discharge signal.

6 DARWIN & CONSCIOUSNESS IN MACHINE CONSCIOUSNESS

In conclusion, we have shown how, through an evolutionary approach to the attention architecture in the brain, we can build up a picture of how consciousness itself evolved, and so begin to justify Darwin's claims which he developed more fully in [1]. This justification clearly needs far more work to fill in the many gaps, in particular the nature of possible conscious experience possessed by the numerous animals known to possess some form of attention control. It could be said that even possession of a form of extended ballistic attention control could lead to functional consciousness: the activity of those attended stimuli reaching their relevant STM would be available to numerous higher level modules that could be conjectured to be able to carry out such functions as thinking, reasoning, etc. However such consciousness could not be claimed to possess any inner self, so would be without

phenomenological consciousness, unless there was an attached corollary discharge attention control signal.

Earlier it was noted that machine consciousness requires some form of consciousness as seen at the human level. Through our evolutionary approach, we have begun to be able to split consciousness into at least two parts, as would agree with [28] in the two forms, active and phenomenological ('access' and 'phenomenal', in Block's terminology). Thus we can consider the similar possibility for machine consciousness. A lower level of such consciousness could be claimed for the functional form, arising solely in the STM and without any inner self to act as owner of the conscious experience. On the other hand it is possible to make more efficient this conscious experience in an owner-controlled manner by going the CODAM architectural route. In the latter case machine consciousness is expected to be more efficient than the lower level, functional-only, consciousness.

The answer to the title of the paper is thus that consciousness can thereby be divided, along the evolutionary route to CODAM. Without a corollary discharge of the attention movement control signal then there will be no owner of the conscious experience, so there will actually be no such experience per se. However such a lower level of consciousness, essentially that of a zombie, could well enable an animal to survive very effectively in a suitable environment. Such may be the situation of the cows, horses, sheep, cats and dogs mentioned earlier: they may not have a corollary discharge signal so no ability to experience what they are functionally conscious of. But that is an empirical question which can be properly determined, if correct, by careful and extended experimentation.

REFERENCES

[1] Darwin C (1871) *The Descent of Man* London: John Murray

[2] Libet B (1981). The experimental evidence for subjective referral of a sensory experience backwards in time: Reply to P. S. Churchland. *Philosophy of Science*, **48**:181-197.

[3] Zahavi D (2005) *Subjectivity & Selfhood* Cambridge MA: MIT Press

[4] Dennett D (1991) *Consciousness Explained* London UK: Penguin Press

[5] Taylor JG (2007) CODAM: A Model of Attention Leading to the Creation of Consciousness. *Scholarpedia* 2(11):1598

[6] Taylor JG & Fragopanagos N (2011) Attention and Control: The New Networks *Cognitive Computation* (to appear)

[7] Fox JJ & Simpson GV (2002) Flow of activation from V1 to frontal cortex in humans *Experimental Brain Research* 142:139-150

[8] J. G. Taylor (2010d) A Neural Model of the Loss of Self in Schizophrenia *Schizophrenia Bulletin* (on-line: April 23rd, 2010)

[9] M. Corbetta, G. Patel & G. Shulman G (2008) The Reorienting System of the Human Brain: From Environment to Theory of Mind. *Neuron* 58:306-324

[10] G. G. Gregoriou, S. J. Gotts, H. Zhou and R. Desimone (2009) High-Frequency, Long-Range Coupling Between Prefrontal and Visual Cortex During Attention *Science* **324**:1207-1210

[11] S. Bressler, W. Tang, C. M. Sylvester, G. L. Shulman and M. Corbetta (2008) Top-Down Control of Human Visual Cortex by Frontal and parietal Cortex in Anticipatory Visual Spatial Attention *J Neuroscience* **28**(40):10056-10061

[12] Xu Y & Chun M-M (2006) Dissociable neural mechanisms supporting visual short-term memory *Nature* 440:91-5

[13] M. Desmurget & S. Grafton, "Forward modeling allows feedback control for fast reaching movements," *Trends in Cognitive Sciences* vol. 4(11) pp. 423-431, 2000

[14] J. G. Taylor (2010c) The I's Eye View of Its Consciousness *Journal of Consciousness Studies* 17(1/2):95-117

[15] L. A. Sass & J. Parnas (2003) Schizophrenia, Consciousness and the Self. *Schizophrenia Bulletin* 29(3):427-444

[16] R. Sokolowski. *Introduction to Phenomenology*. Cambridge: Cambridge University Press; 2000.

[17] J. G. Taylor. (2002). Paying attention to consciousness *Trends in Cognitive Sciences*, 6 (5), 206-210.

[18] J. G. Taylor (2002) From Matter to Mind *Journal of Consciousness Studies* 6:3-22

[19] Hopf J-M, Luck SJ, Girelloi M, Tillman H, Mangun GR, Scheich H & Heinze H-J (2000) Neural Sources of Focussed Attention in Visual Search *Cerebral Cortex* 10:1233-1241

[20] Van der Stigchel S, Heslenfeld DJ and Theeuwes J (2006) An ERP study of preparatory and inhibitory mechanisms in a cued saccade task *Brain Research* 1105:32-45

[21] Sergent C, Baillet S & Dehaene S (2005) Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience* 8:1391-1400

[22] Merriam EP, Genovese CR & Colby CL (2007) Remapping in Human Visual Cortex *J Neurophysiology* 97:1738-1755

[23] P. Cavanagh, A. R. Hunt, A. Afraz & M. Rolfs (2010) Visual Stability based on remapping of attention pointers. *Trends in Cognitive Sciences* (in press)

[24] N. Robitaille and P. Jolicoeur (2006) Fundamental Properties of the N2pc as an Index of Spatial Attention: Effects of Masking *Canadian Journal of Experimental Psychology* **60**(2):101-111

[25] M. Kiss, J. Van Velzen and M. Eimer (2008) The N2pc component and its links to attention shifts and spatially selective visual processing. *Psychophysiology*, **45:**240-249

[26] S. J. Luck and S. A. Hillyard (!994) Spatial filtering during visual search. Evidence from human electrophysiology J *Experimental Psychology: Human Perception and Performance* **20**:1000-1014

[27] N. Robitaille, S. Grimault & P. Jolicoeur (2009) Bilateral parietal and contralateral responses during maintenance of unilaterally encoded objects in visual short-term memory: Evidence from magneto-encephalography *Psychophysiology* **46**:1090-1099

[28] Block N (1995) On a confusion about a function of consciousness, in Block, N., Flanagan, O. and Güzeldere, G. (eds.) *The Nature of Consciousness*, Cambridge, Mass.: MIT Press.

Would a super-intelligent AI necessarily be (super-)conscious?

Steve Torrance
Centre for Research in Cognitive Science (COGS),
University of Sussex, Falmer, Brighton BN1 9QJ
Email: sbtorrance@hotmail.com

Abstract. Some AI futurists are predicting a 'technological singularity' when artificial super-intelligence, or 'AI++', as Chalmers has called it, [1] explodes onto the scene. There are many reasons to be sceptical about the 'explosionist' scenario. Yet the notion of artificial super-intelligence appears to be relatively intelligible, as a projection from current the current AI state-of-the-art, and as a gradually emergent process, if not as a singularity.

In this paper the explosionist position and the notion of superintelligence are considered, in the context of MC research. Suppose AI++ were to come about: would real MC have then also arrived, 'for free'? Does the idea tempt you, as an MC investigator?

A number of interconnected questions are addressed: What are the various positions that might be adopted on the issue of whether an AI++ would necessarily (or with strong likelihood) be a *conscious* AI++? Would a conscious super-intelligence also be a *super*-consciousness? (Indeed, what meaning might be attached to the notion of 'super-consciousness'?) What ethical and social consequences might be drawn from the idea of conscious super-AIs or from that of artificial super-consciousness? And what implications does this issue have for technical progress on MC in a pre-AI++ world?

1 INTRODUCTION

Certain tech-enthusiast circles have become increasingly prone to proclaim the coming of the 'Technological Singularity' or 'Superintelligence Explosion', an event where artificial intelligence will come to match and then rapidly overtake levels of human intelligence [1-8]. The Singularity Institute for Artificial Intelligence, based in San Francisco, welcome this event and indeed seek to hasten it. Superintelligences will, they believe, solve global problems like cancer, AIDS, world hunger, etc. [9] The fact that such a development might be one which should be discussed democratically across world communities, to be fair, does not seem to have great weight with the members of this institute, or at least it has not dampened their social-engineering zeal [10].

Leaving aside the ethics of going for broke on the singularity in such a single-minded way, what reasons can be given for its likelihood, or even its imminence? The dynamic of the development of super-intelligence is usually seen as being based on the combination of a number of factors. Two major such factors include: (a) the ability of machines to take a progressively more active role in their own improved design; and (b) the inherent tendency for key technology measures (e.g. number of transistors on a given surface of silicon, processing speed, lowering of cost in production of components, etc.) to double every 18 months to two years – variants of Moore's Law [11]; [5]. It is argued (see [1], for example) that such factors, taken together, may produce an eventual explosion in super-AIs. However, even without such an 'explosionist' scenario, there is a strong possibility that future AI-technologies may incrementally progress over many decades or centuries from the rather limited levels of intelligence broadly shown today, to levels which far surpass human intelligence at some deep future time.

In his recent extended study of the Singularity, Chalmers uses the term 'AI+' to refer to agents whose intelligence exceeds that of average humans by just a little, and 'AI++' to refer to agents whose intelligence levels are to human levels rather as human intelligence is to rodent intelligence [1]. (I use the terms 'super-intelligence', 'super-AI' and Chalmers' term 'AI++' more or less interchangeably in what follows.) Chalmers claims that one may reasonably predict that AI+ may be achieved at some point in the not-too-distant future. He further predicts that once AI+ is obtained, AI++ should occur not too long afterwards. One possible argument in favour of the second prediction is this: if AI+ has been successfully produced, then it is plausible that such AI+ agents should be able to produce improved versions of themselves – after all, they were produced by beings less intelligent than them, namely humans (and other AIs). A process of recursive self-improvement would thus seem to be a reasonable scenario. Moreover, if some Moore's Law-style technological acceleration were to operate, then each successive wave of self-improvement would take a shorter time than the previous one, so that the emergence of AI++ might occur in a relatively brief, indeed explosive, period of objective time after the appearance of AI+. Even without such an explosive time-scale, the necessary recursive self-improvement process may take place on a longer time-scale, more amenable to human observation and control, perhaps.

2 THE DROP-OUT QUESTION (DOQ)

It is tempting to dismiss the singularity or explosionist scenarios as entirely baseless. But it would be ill-advised simply to turn one's back on the subject. In view of the fundamental change in human history (perhaps indeed in evolutionary history) that a super-intelligence explosion might bring about, were it indeed to occur, we need to think as clearly as possible about its implications, in the event of its actuality. Also, quite apart from its far-reaching social impacts, the possibility of AI++ has important theoretical ramifications. Indeed, as shall be argued that, from the point of view of the Machine Consciousness (MC)

community, the possible advent of super-AI brings into profile a number of intriguing issues about the status of MC. [1]

One such issue may be called the 'Drop-Out Question' (DOQ). Suppose one or more super-intelligences were to come into existence roughly as predicted by Singularity apostles. (For reasons given below, a lot of work may have to be done in specifying certain details in the description. For example: How generalized in scope would their intelligence be? How narrow or wide a conception of intelligence would be exemplified by such superintelligent beings? Would the intelligence be accompanied by other mental properties, particularly consciousness? And so on. But for now, let us leave such details undetermined except where needed to progress the discussion.)

What about the consciousness of super-AIs? One possibility is that super-AIs would necessarily be conscious beings. An alternative possibility is that they would be super-smart, environmentally adaptive, highly powerful, zombies – that is, cognitively high-functioning, but totally without conscious awareness (in the sense of 'phenomenal consciousness', at least). A key question is: Would consciousness simply drop out of the development of AI++ 'for free', as it were, as a kind of natural concomitant of the comprehensive high-grade cognitive capacities of such agents? Alternatively, would there have to be a particular kind of design constraint built into the specification of super-AIs to ensure that such beings were not just super-smart but also had (at least) the same degree of self-awareness or 'what-it-is-likeness' that humans have? Or would it indeed be the case – as many AI-sceptics have claimed – that no purely AI project could generate genuine qualitative or phenomenal awareness (as opposed to behaviour which simulates such awareness), if based simply upon a progression from present-day computational hardware and software technologies?

(A caveat. Of course the progression to super-intelligence may involve the development of some quite novel technologies completely unknown to us at present, or it may essentially involve major developments in technologies that are currently known but that exist today only in highly embryonic forms – such as biocomputing, quantum computing, synthetic biology, and so on. We will concentrate on the possibility that the production of super-intelligence simply develops on current computational and robotic techniques, but in ways that are vastly improve on current achievements in terms of performance, efficiency, design, etc.)

For now let us focus on the possibility that super-intelligence may automatically bring consciousness 'for free', as it were. How might we evaluate such a possibility? The DOQ, it is suggested, provides the opportunity to reframe certain important debates within the MC research community. Many of the key research themes of the present and of other MC discussion forums – such as information integration, self-models, neural architectures, embodiment, imagination, and so on – can be discussed in the context of the DOQ, not in terms of the rather limited functionalities of currently achieved bench models, but rather in terms of hypothesized ultra-high-performance systems which, one assumes, would have all or most of the cognitive features of human intelligence (plus maybe some more), and whose levels of attainment in those respects far outclassed the best human levels.

The DOQ also highlights another important issue concerning Machine Consciousness as a research area. Is MC, as an experimental and theoretical study, something that can be considered to be a sub-department of mainstream AI? Or is it something which carries a special set of performance-criteria or success-criteria, in that it targets, not (just) intelligence, as is the case with AI, but also consciousness?

3 RESPONSES TO THE DOQ

There many different possible responses to the DOQ. Here are six.

a. Hard scepticism: consciousness does not drop out from super-AI because a comprehensive super-intelligence is theoretically impossible using current day AI techniques. So the question of consciousness emerging from super-AIs doesn't arise.
b. Soft scepticism: it may be possible to produce a highly optimized computational agent displaying super-intelligent performance, but such an agent will be likely to be completely non-conscious, whatever its performance characteristics.
c. Only functional consciousness can be replicated: a super-AI will have many of the features of functional consciousness, but none of that implies phenomenal self-awareness, which is what really matters about consciousness.
d. Novel technologies are needed: if a super-AI is built using only highly optimized versions of current computational technologies, then no consciousness could result. To achieve artificial consciousness, as opposed to artificial cognition, quite different technologies are required, such as bio-synthesis.
e. A Chalmers-style conclusion: Phenomenal consciousness contingently arises if the right cognitive properties are present, which may well be present in a super-AI. So a super-intelligence is quite likely to be phenomenally conscious.
f. A Dennettian conclusion – no sense can be made of phenomenal consciousness as a separate kind of mental feature. A super-AI would be likely to have all the functional features of conscious awareness that we have, and the functional features are all there are to consciousness.

It is likely that people who are sceptical about the likely success of the super-intelligence project will also give a sceptical response to the DOQ. The converse, however, is not necessarily true. One may feel optimistic that, *given* the hypothetical arrival of AI++, such super-intelligent agents would instantiate artificial consciousness, while being far from confident about the likely emergence of AI++.

Option (c) – that super-AIs would instantiate functional, but not phenomenal, consciousness – is likely to be favoured by someone who sees the development of comprehensive intelligence of a kind that could assist the emergence of AI+ or AI++ as necessarily going hand in hand with certain functional features of consciousness, such as the development of a unified self-model [12-14], or of a self-aware global workspace system

[1] Singularity enthusiasts would perhaps be seen by a majority of scientists and the general public (to the extent that they are aware of them) as a bunch of messianic, glassy-eyed techies. It might be reflected, with not a little irony, that a similar view might well be taken, from those same perspectives, of the MC community, especially those seeking to create MC instantiations rather than just models!

for phenomenal awareness or conscious deliberation [15-18]. Many MC researchers are thus perhaps likely to agree that progress towards AI+ and AI++ is best served by pursuing cognitive models that incorporate the best results from MC.

The general thrust of much MC research may thus suggest the following broad response to the DOQ. The development of intelligence in humans and other animals is closely associated with the parallel emergence of sophisticated forms of awareness. These forms of awareness are perhaps to be explained in terms of self-monitoring systems, or the possession of rich self-models, or in terms of abilities for offline imaginative scenario exploration, or information integration, highly optimized attention mechanisms, and so on. On this view, rather than consciousness dropping out 'for free' from a super-intelligence system, the development of robust and realistic models of intelligence of the sort needed to achieve AI++ performance levels would in turn have to be informed by robust and realistic models of consciousness. This would be so, it might be argued, because (toy systems apart) consciousness and intelligence must necessarily go hand in hand. So rather than being able to wait for super-intelligence engineers to do their work in order to see the development of artificial consciousness, consciousness engineers and super-intelligence engineers would really need to work hard alongside each other in a cooperative enterprise. This, then, might be one argument advanced by some MC researchers, for favouring a bullish attitude to the DOQ.

4 FUNCTIONALITY, PHENOMENALITY AND THE DOQ

The DOQ is made more complicated by the distinction, frequently made, between phenomenal and functional consciousness. Not everyone accepts that this distinction is valid – Dennett, for example, denies that there is any notion of phenomenal consciousness which cannot be fully understood in cognitive terms [19]. However, many workers in MC have been more circumspect, at least allowing that there is a theoretical distinction between the two (see [20], for example), but arguing that, a sufficiently rich model of functional consciousness would automatically bring phenomenal consciousness in its wake.

To accept such a functional/phenomenal distinction may be conceptually hazardous, however. Recognizing the distinction seems to imply that it would be possible, at least in principle, to have a system, or agent, which was functionally consciousness but not phenomenally consciousness. But such a possibility may be as incoherent as thinking that the Cheshire Cat's grin could persist when the rest of the cat had disappeared. It could be argued that a system or agent that was 'only functionally conscious' (and not phenomenally so) would not be a conscious system or agent at all. Thus, it might be said that, without the 'what-it's-like' of phenomenality there would be no experiential awareness, and hence no reason to consider that consciousness was present at all in the agent. It should be made clear that such a position could accept that a (phenomenally) non-conscious system could have much (maybe even all) *of the functionality of consciousness*, while still being non-conscious *per se*. So this view would still accept that there are functional features to consciousness on the one hand and phenomenal features on the other, and that the former may exist in a system without the latter, while still asserting that a system in which *only* the functional features were present would not be a conscious system as such.

To distinguish in this way between having functional consciousness and having the functionality, or the functional features, of consciousness may seem to some like splitting hairs, but to others it may seem all-important. At the very least, it seems to have important *ethical* consequences. Many people would agree that it is the phenomenal features of consciousness rather than the functional ones that matter ethically. We would, or should, care about how we treated a system with phenomenal features of consciousness, or even, perhaps, whether to bring it into existence ([12,13, 21]) but there is no such obligation to care about how we treated a system with only functional features – so the argument would go.

We will discuss the ethical aspects of MC below, but for now let us note that bringing in this ethical consideration imparts a keener edge to the DOQ question. In asking whether (phenomenal) consciousness will simply arrive 'for free' with a super-intelligent agent we are, it may be insisted, asking whether a super-intelligent agent would have the kinds of properties of consciousness that were ethically relevant to how we were obliged to treat it (and perhaps how it ought to treat us).

To see how complex and subtle the relationship is between functional and phenomenal aspects of consciousness, consider a concrete hypothetical case: an imaginary artificial system which reproduces the functionality of the McGurk effect [22]. If people hear a recording of someone saying 'big' and at the same time a video of some lips mouthing 'gig', they are likely to report that they heard (and saw) 'dig'. It should not be too difficult to build an artificial system that replicates that response. In the case of the human undergoing the McGurk effect, there is not just a third-person observable response but a first-person phenomenal experience which is being supported. What about the artificial system?

In order for phenomenal consciousness to be attributable to such a system, it is surely not sufficient for it to reproduce a specific human discriminatory behaviour that evidences the occurrence of certain 'qualia' when that behaviour occurs in the human subject: some much more complex global or architectural conditions must obtain. One of the key goals of MC research is, presumably, to try to work out what these global, architectural conditions for phenomenal consciousness are, rather than just to reproduce relatively specific, isolated instances of 'qualia' such as those experienced by people subject to the McGurk effect. As Tononi has remarked, 'Phenomenologically, every experience is an integrated whole...' [23]. So the human subject's phenomenal experience of the 'dig' syllable is a small chunk of the overall experiential manifold that the subject is undergoing at that moment. It is the integrated nature of this experiential manifold that gives it the phenomenological quality that it has (maybe in conjunction with other global features). A dedicated artificial phoneme recognition system has no complex experiential manifold: it only registers a limited set of speech sounds. It is the overall cognitive or information architecture of the human experiential system that enables it to be a centre of phenomenal experience. (Perhaps some defenders of the distinctiveness of phenomenal consciousness will want more – but at least this should be admitted by them, surely.)

Summarizing the points from the above discussion, any answer to the DOQ, should, we suggest, take cognizance of the following:

A. The possibly tempting position that a super-intelligent agent is bound to have be functionally conscious, *even if not phenomenally conscious*, suffers from the difficulty that functional-consciousness-minus-phenomenal-consciousness may not be consciousness at all, since it is the phenomenality that makes it conscious *per se*;

B. One reason why it is important to drive this conceptual wedge between functionality and phenomenality in this context is that there seems to be an ethical significance to states of phenomenal consciousness that are apparently missing in the case of merely functional consciousness (discussions about artificial consciousness seem to have an ethical 'bite' or 'piquancy' about them that discussions about artificial intelligence lack, or that the latter possess, perhaps, only if questions about consciousness are thought of as downstream of the discussion);

C. Almost certainly, the conditions of phenomenality won't be fulfilled simply by reproducing this or that phenomenal or qualia-type state in isolation (as in the McGurk effect) but by reproducing some broad, global, architectural conditions. These conditions may be broad cognitive-architectural features (for example a global workspace architecture [15], or a cognitive structure supporting something like Metzinger's transparent phenomenal self-model [12]). But equally, they may have more to do with deep biological, or metabolic, features of the system – for example, whether the system is sufficiently like a living organism to be dependent upon, or constituted by, a dynamically self-maintaining network of processes existing in a far-from-equilibrium state. No one can pretend that the debate between these two broad options is closed currently; and if the latter viewpoint is correct – i.e. if the phenomenality of consciousness emerges from these deep-seated bio-functional properties rather than merely from cognitive-functional properties, then there seems to be little ground for assuming that an AI++ would, merely from having a rich set of intelligence-related performance characteristics, also develop functional consciousness.

5 INTELLIGENCE, SUPER AND NOT-SO SUPER

Before one can really assess claims about the emergence of super-intelligence, or AI++, there are, in any case, a number of issues to do with the nature of plain old 'intelligence', as we apply that term to humans, and possibly to mice. In a discussion on machine consciousness one does not really want to deviate too much into core issues concerning AI, but in the present context some discussion is necessary. As we will see, it is of fundamental importance to how we view the likelihood of progress in the field of MC.

A strong assumption underlying some of the literature on the singularity or the intelligence explosion, is that human and machine levels of intelligence can be measured on a single scale – and indeed that the intelligence of human and various kinds of non-human animals can also be graded in an unproblematic way. This may be so, but it can't simply be taken as read. For one thing, all forms of natural intelligence are found in creatures which have a long and interlinked evolutionary history, and in which the various skills that are clustered under the term 'intelligence' developed in a relatively integrated manner as those creatures coped with various kinds of environment. Intelligence in machines, as it has been studied and modelled in AI over the last six or so decades, has not emerged as a result of a natural evolutionary process (and the kinds of artificial evolution found in A-Life models are, on the whole, relatively simplistic and abstract). This is one evident ground for asserting a strong discontinuity between naturally occurring intelligence-features and 'intelligence' displayed in machine models and agents.

On the other hand there are clearly some important evolutionary discontinuities or specificities in humans – language, society, culture, history – that would suggest a strong demarcation between what might be called 'intelligence' in humans and what might qualify for that title in other natural creatures. This makes the idea of any smooth scale of comparison in 'intelligence' from (e.g.) rodent to human to machine rather problematic.

Both these discountinuities seem to be disregarded by writers on the singularity. Chalmers, for example, defines 'AI++' in a way that explicitly appeals to such a continuous scale [1]. Of course some broad comparisons can be made: mice can solve maze problems but can't do calculus, and there are no doubt many 'cognitive' feats that future computational systems may perform which are beyond human capability. But something important distinguishes the intelligence of both a human and of a mouse from that of an AI system, at least current AI systems. Whereas the intelligence of both humans and mice allow them to be pretty-well self-standing in their respective niches, the intelligence of any present-day artificial cognitive agent does not free that agent from being fundamentally dependent upon human support for its continued existence. Of course this may change if the predicted progression to AI+, and onwards to AI++, occurs (perhaps a necessary criterion of reaching AI+, let alone AI++, is indeed that any agent qualifying for such a title must be more or less self-sustaining in this sense). But there seems to be no obvious *guarantee* that such an independent self-standing status will be developed.[2]

This relates to another important feature of artificial forms of intelligence: the fact that current AIs are usually designed to perform a specific set of tasks, and have little or no abilities beyond the boundaries of those tasks. Much work is being done on developing models of 'artificial general intelligence' (AGI), where, instead of 'islands' of domain-specific ability, such systems will exhibit mainlands of operative capacity [4, 25-27]. Significant, and interesting, approaches to AGI are being developed by some workers who also strongly identify with the Machine Consciousness research programme. For example, Stan Franklin has developed, with Bernard Baars and colleagues, LIDA, a prototype system for modelling general decision-making [15,25]. LIDA's design revolves around the insight that the deliberations of any naturally intelligent decision-maker must draw upon the resources of that creature's consciousness. So a good model of general decision-making must, on this approach, also incorporate a good model of consciousness (in LIDA's case the Baars Global Workspace model). Indeed it may well be the case that much other work in Machine Consciousness will help to progress work in AGI, because of the clear connections between functional consciousness and generalized intelligent cognitive capacities.

[2] If an AI+ or AI++ were to be self-sustaining in this sense, it perhaps might follow that the latter had its own, intrinsic goals or teleology. This might, in turn, be considered a strong ground for saying that it possessed phenomenal consciousness. But a lot needs to be done to make this argument water-tight.

Nevertheless the development of really effective generalized intelligent decision-makers or agents – whether functionally conscious or no – may be fraught with difficulties. It seems likely that any agent that might qualify as an AI++ has to be built around a robust AGI model. If that were so, the likelihood of the superintelligence explosion (or its emergence at a more leisurely pace) would appear to be crucially dependent upon the success of AGI research. Yet, prototype models aside, no one has come near to developing anything that could be taken as a serious practical contender for a true AGI at present.

Moreover, how general need an AGI be to be an AGI? One could imagine a system being developed that had a broad set of 'intellectual' capacities, so that it performed faultlessly and seamlessly in many contexts, but which still had serious gaps when compared with certain aspects of human performance (or even canine or corvid performance). One might even have a system that appeared, when judged against a whole raft of criteria, to clearly rank as an AI++, but which, in a minority of respects, perhaps, failed miserably. Some of these areas of underperformance may be pretty obvious and therefore discountable or remediable (compare a Nobel laureate's ineptitude at practical tasks like driving a car or buying groceries). But some deficits might be subtle and hard to spot – and yet may vitiate the judgment or the wisdom of the AI++ agent in ways that could have crucial consequences for human destiny.

This leads to a wider question – what kinds of aptitudes count as 'intelligence'? The field of intelligence-measurement has been plagued by the street-lamp syndrome. Like the drunk looking for a lost set of keys under a light because it was easier to see there, psychometricians concentrate on easily-measurable features of cognitive fitness, such as deductive or verbal reasoning, focused attention, visuo-spatial working memory, etc., while paying relatively little heed to what may be crucial, but less operationally amenable, features such as appropriate emotional response, physical manipulation skills, creative aptitudes, and so on. The kind of bias that is built into the term 'intelligence' by frequent usage makes people naturally think of skill in chess or mathematical problem-solving; yet the sort of embodied, empathetic intelligence that is necessary to handle and nurture a new-born baby may draw upon quite different kinds of intelligence – ones which are far less frequently studied or measured.

An AI++ may, by definition, possess an abundance of skills of the first sort, but yet may have dangerously few skills of the latter sort. The notion of intelligence is, plausibly, a cluster notion with many different facets. [27] Certainly it would be dangerous if an AI++ built on a highly restricted notion of intelligence were to be given (or were to take upon itself) the responsibility to address the intractable problems of human society that humans themselves have been so poor at solving (yet it is part of the rhetoric of Singularity apostles that super-AIs are could serve as saviours of humanity in just this way [9]).

For all these reasons, and others, the notion of super-intelligence is far from straightforward. There may thus be perilous hidden reefs along the channels which lead to the development of AI+ and AI++. Such developments may fail because of the inherently vague, multivocal, and inherently open-ended nature of the notion of 'intelligence'; and worse, there may be no clearly assignable success-conditions to the task of achieving artificial 'super-intelligence'. Nevertheless, while showing that predictions about AI++ need to be handled with care, they do not show that AI++ is an impossibility – so the DOQ still remains a key problem worth considering.

6 'SUPER-CONSCIOUSNESS'?

How does thinking about AI++ help us to hypothesize about the kinds of artificial consciousness that might emerge? Let us suppose that some form of consciousness (and consciousness, indeed, of a phenomenal kind) were to be present in an emerging super-AI agent. What kind of consciousness might this be? Would it be 'just like' our consciousness in most major respects? Or would it be appropriate to say, by analogy with Chalmers' characterization of AI++, that it was as different from human consciousness as the latter is from rodent consciousness? Would a phenomenally conscious super-AI have the same kinds of 'feels' as a smart human or would the large disparity in 'intelligence level' mean that the conscious states of the super-AI system would exhibit some kind of quantitative, or qualitative, difference from human conscious states? (For that matter, do the phenomenal states of a human whose cognitive capacities are situated towards the high end of the intelligence distribution differ either quantitatively or qualitatively from those of a human with scores at the low end?)

It is thus tempting to talk in terms of 'super-consciousness' or 'super-experience' as a state concomitant to super-intelligences. But what might such a notion encompass? Would a super-conscious being feel pleasures and pains more keenly? Would it have the capacity for a more vivid range of sensory experiences? Would such a being be prone to being rocked by far more tempestuous emotions than even the most highly-strung human? Or would its highly developed intelligence enable it to master emotional forces to a greater degree than humans can generally do? Or, by contrast, could it be that the kind of AI++ that is most likely to emerge is one which has highly developed cognitive powers, but no emotional capacity whatsoever?

There has been some discussion of the idea that consciousness may come in graded levels or quantities, within the MC literature. Perhaps the most general such measure is Tononi's Φ, which equates the degree of consciousness in a system with a combination of integration and differentiation in the information held by the system [23]. Tononi is explicit that artificial systems can have Φ measures in the same way as biological organisms can [28], but has not discussed, in detail, whether any future super-intelligent agent would be likely to have measures of information integration in excess of human levels in proportion to the degree that their intelligence levels exceed human intelligence levels. Seth, who has proposed a measure of degrees of consciousness in terms of causal integration, has argued that the Φ measure can be trivialized, as it could be found in arbitrarily high amounts in relatively simple systems such as fully connected Hopfield nets [29].

The presentation of Tononi's Φ, and other such measures of consciousness , has mainly taken an anthropocentric stance, the interest being focused on how to find a systematic way to compare human consciousness with that of other biological organisms; and by extension how to provide a rigorous way to specify the degree to which different machine models or instantiations of consciousness may approach human levels.

Arrabales and colleagues have produced a highly articulated and detailed metric, ConsScale (version 2), which perhaps provides the most useful elaboration of the multiple facets that

must come in to a full comparison of levels of consciousness across the phylogenetic scale [30]; see also [31]. Arrabales *et al*'s Quantitative Score (CQS), calibrated to run, in an exponential fashion, from 0 to 1000, has what is termed a 'super-consciousness' measure at its maximum end, but this is described in rather tentative terms (being characterized in terms of ability to manage multiple streams of consciousness), and does not address how the CQS might be adapted to measure levels of consciousness and associated capacities in agents with the far greater cognitive powers envisaged in Singularity discussions.[3]

Clearly there is some interesting future work to be done in investigating how far these approaches to measuring or grading levels of consciousness (or competing approaches) may be explicitly coupled with the AI+ and AI++ scenarios. We have shown that any such speculations need careful qualification and discussion in the light of difficulties we have raised concerning both the intelligence and the phenomenological aspects. Nevertheless there is a real possibility that super-AI agents may come to be seen as having enhanced or deepened phenomenal or affective capacities which somewhat match its super-human intellectual capacities. (Indeed, such super-AIs, no doubt being party to the debate, may elegantly and persuasively insist on their own super-conscious levels.)

If indeed this is so, then a number of uncomfortable moral issues come to the fore. Our present-day ethical systems are currently based upon the assumption that every human being is entitled to similar consideration in terms of rights to avoid suffering and to seek personal satisfaction and fulfilment. Such equality of consideration across the human race is built into moral documents such as the 1948 Universal Declaration of Human Rights, the 1950 European Convention on Human Rights, and so on. Humanity has generally taken it for granted that such rights do not apply to non-human organisms, despite the fact that degrees of consciousness are often readily attributed to many other biological species. As a moral community, humans do often concede that other creatures have some level of ethical status, and they do so (at least in part) because of their possession of cognitive and phenomenological capacities in varying degrees. Because of the variations a clear moral hierarchy is generally taken to operate between humans and other creatures, even though there is little consensus about the detailed structure of such a hierarchy. .

So with the advent of AI++ agents it seems that we should, in fairness, expect new, higher, layers to be added to that moral hierarchy, especially if such beings possess, not just super-human levels of intelligence, but also levels of phenomenal experience that far exceed human experiential capacities? Would it even have to be the case that the normative principles inherent in the best human ethical systems commit humans to conceding that the moral interests of super-AIs should *take precedence* over the moral claims of humans?[4] It is difficult to see how such a moral hierarchy extending beyond humans could be avoided if we are to be consistent. And remember, this may well not just be a 'private' debate by humanity *entre nous*.

This is a bleak conclusion. Humans have been used to thinking of themselves as at the top of the pile. But it seems to be an implication of the super-intelligence thought-experiment, at least if some kind of super-consciousness were to emerge alongside the super-intelligence. But surely there would then, in turn, be overwhelming reasons to prevent the emergence of super-AIs in practice. Far from such super-AIs being able to solve the ills of humanity, super-AIs may well feel justified in subjugating humans in just the ways that we humans have, for millennia, subjugated other, less intelligent animal species. Moreover, consistency with the dominant human practice of subjugating less intelligent animals would require humanity to approve of our being treated in this discriminatory way (or else we had better all become vegetarians fast!)

7 CONCLUSION: CONSCIOUSNESS, ETHICS AND SUPER-AI

There are thus some tough, and deep, issues underlying the technical work in MC research. We have used the context of current explorations of the idea of a super-intelligence explosion not because we are arguing that such an explosion is imminent, but because it helps to dramatize some of these deep issues, and to clarify some of the options. What emerges from such an examination is that there are several important connections between intelligence, consciousness and ethics, in the context of AI, past and future.

Where does Machine Consciousness stand in relation to the heritage of AI and cognitive science as the latter has developed over the past half-century or so? Our discussion implies that MC research has to be seen as lying at the heart of that heritage, even though there may be a temptation for some MC workers to see what they do as being somewhat detached and orthogonal to mainstream AI research. We would argue that the global, architectural models of functional consciousness currently under development in the MC community are likely to play a vital role in the construction of robust, future large-scale models of cognition or of AGI. In so far as such developments in comprehensive or generalized AI are central to the hypothesized emergence of super-intelligence, the development of richer artificial consciousness models will play a central role in the progression towards such super-intelligence (whatever we might feel about such an outcome). This picture stands in contrast to the one that might be summoned up on first considering the DOQ, namely a scenario in which, during the progression towards super-intelligence, somehow, mysteriously and passively, consciousness simply arrives 'for free' within the package.

Further, as we have seen, the prospect of a world of super-smart AI agents clearly raises a number of important ethical issues – How will/should we treat them, or they us? What kind of mental properties do we need to assume are present before our treatment of them becomes ethically relevant? How do we detect the presence of such properties? Further, is there an ethical hierarchy of treatment or consideration which could include humans, various kinds of animals, and various sorts of artificially conscious agents, including ones that far transcend current human levels?

[3] Metzinger [13] has engaged in some imaginative projections on how future artificial intelligences may transcend current humanity in phenomenological, and moral, as well as cognitive respects. The present discussion is very much indebted to Metzinger's treatment, but we would differ from his approach in key respects.

[4] We do not here raise the even more vexed issue that the super-intelligences which may be thought to be given moral precedence over (normal) humans, might possibly also include cognitively supersized, or uploaded, or otherwise enhanced, *humans*.

Leaving aside the specific ethical and social impact issues of a potential super-AI era, there are clearly some points to be made concerning the relation between the study of consciousness in both human and artificial agents, and moral questions concerning such agents. As already observed, questions concerning machine consciousness appear to have an ethical 'bite' in a way that questions concerning 'mere' AI systems do not. Consciousness – at least of the phenomenal sort – 'matters' in a way that merely functional aspects of mind do not. One might say (in a way that perhaps requires careful qualification) that questions concerning the attribution of consciousness have an ethical status that other kinds of psychological attribution lack. [32, 33]

Perhaps this will also help to explain, at least in part, why it is, as some have claimed, that MC is very closely associated with the field of artificial, or machine, ethics. [34-36] Clearly, claims about the conscious states of artificial agents seem to raise other claims to do with how such agents should be treated – that is, about their status as moral *patients* or *recipients*. Metzinger has suggested that machine consciousness, when properly instantiated, would introduce the possibility of suffering in the machine agents to which that property applied. He has gone so far as to say that, in virtue of that possibility, and of the resulting circumstance of our bringing unnecessary suffering into the world, the research activity of MC should perhaps be banned, and certainly considered in a morally precautionary light [12, 13].

There is another important reason why MC and Machine Ethics have a close association – why they are, as Wallach, Allen and Franklin put it, 'joined at the hip' [36]. Consider ethical agents, or 'producers', in the sense of actors who make morally significant decisions and who are deemed to have moral responsibilities (in contrast to ethical 'consumers' or recipients, who are the targets of such morally significant action). [34] If super-AIs are to be the kinds of ethical producers that we should take seriously in a moral sense, and that we should want to accept as full members of our moral community, they surely also have to be conscious agents in a rich sense. Arguably, because of the central role played by consciousness in determining what counts as morally significant for ethical action, any ethical agent must surely understand consciousness 'from the inside', as it were – that is, as a being who is conscious in the ethically engaging sense.

Wallach *et al.* make a specific claim [36] that any system that fits the bill of being a moral agent in this sense will also qualify as a conscious agent in having the appropriate kind of MC cognitive architecture. Suppose that eventually AI++ agents do come about, and further, that the cognitive architecture that they exemplify conforms reasonably closely to, say, the LIDA/GW pattern described by Franklin and others. Then if the above claim is correct, such AI++ agents will at least have the functionality of conscious creatures. We questioned earlier that whether any such system, merely in terms of its computationally specified cognitive architecture, will have other than functional consciousness (or, as we put it, the functionality, rather than the phenomenality, of consciousness). But this is clearly an open question at present – nothing said here has been intended to foreclose it. So there is indeed force to the argument that an AI agent with all the functionality of consciousness will be able to function as a moral 'producer', in the sense outlined just now.

Clearly, one must be cautious, for two reasons. First, suppose it is correct to claim (as we earlier intimated) that if a being is phenomenally conscious it has genuine moral interests (is a genuine moral 'recipient'); and that if it is not phenomenally conscious it does not. Then there are real issues, to do with false positives and false negatives, affecting how we treat artificial agents that we consider to be phenomenally conscious. If we cede moral interests to super-AIs who are, *in fact,* non-conscious in the phenomenal sense, we may be wasting valuable resources on beings that have no moral need for them. Conversely, if our belief that they are non-conscious, phenomenally speaking, is *mistaken*, then we would be doing them a great moral injustice by refusing to grant their needs (assuming we had the power to do so). So attributing consciousness to AI agents in a way that has ethical 'bite', in itself carries great (moral) responsibilities. (But see [40, 41] for a contrasting view of how consciousness-attributions relate (or fail to relate) to how we think about moral interests in the case of both humans and of future AI agents.)

Second, in our conception of moral agents or producers mentioned above, we considered the view that consciousness (at least of a functional kind), and the capacity for moral decision-making, go together. We would surely hope that any super-AI agents that we brought into existence (or that we seeded for other artificial agents to bring into existence) would deliberate in ways that were responsible, just, benevolent, and so on – that is, in ways that would be 'friendly' to humans. [3, 37-39] Yet, in the generic sense of 'ethical agent' (or 'producer'), villains may be no less ethical agents than saints are: a villain has *the capacity* for benevolence and other kinds of moral virtue, but opts for a different path. (Also, of course, villains are no less conscious than morally upright individuals.) But perhaps moral venality always results from some rational failure, which in turn can be laid at the door of some intelligence deficit that an AI++ can be guaranteed, constitutively, not to have. However, that point would need a lot more debate and reflection.

Surely any project to produce super-intelligence needs to put its money on the idea that increased rationality or general intelligence also brings increased ethical insight in the sense of commitment to the broad interests of conscious beings as such, and as far as is humanly (!) possible, ensuring that such commitment was maximized. The supposition would thus be that super-AIs, were they to come about in a relatively human-controlled way, would be relied upon to understand and participate in our moral discourse, and to agree with our best moral practices, thereby accepting a moral universality that gives due weight to the interests of both humans and artificial conscious agents (not to mention other creatures). But this, of course, is a rather tenuous supposition, and at present we have only the most slender hope that a singularity or super-AI era would not develop in a way that would threaten human interests and be progressively beyond human influence.

Acknowledgements. Thanks to Mike Beaton, Ron Chrisley, Robert Clowes, David Gamez, Murray Shanahan, Wendell Wallach and Blay Whitby for useful discussions on various issues covered in this paper.

References

[1] Chalmers, D J (2010). 'The Singularity: A Philosophical Analysis' *Journal of Consciousness Studies*. **17**(9-10) pp. 7-65.

[2] Bostrom, N (2005) 'A History of Transhumanist Thought.' *Journal of Evolution and Technology*, **14**, pp. 1-25

[3] Yudkowsky, E. (2008). 'Artificial intelligence as a positive and negative factor in global risk', in N.Bostrom and M. Cirkovic (eds). *Global Catastrophic Risks.* Oxford: OUP, 91-119

[4] Goertzel, B (2007) 'Human-level artificial general intelligence and the possibility of a technological singularity' *Artificial Intelligence,* **18**, pp. 1161-1173

[5] Good, I.J. (1965). 'Speculations concerning the first ultraintelligent machine.' In F. L. Alt and M. Rubinoff (eds) *Advances in Computers.* **6**: 31–88, Academic Press.

[6] Kurzweil, R (2005). *The Singularity is Near: When Humans Transcend Biology.* NY: Viking Press.

[7] Sandberg, A (2010) 'An Overview of Models of Technological Singularity' *Proc. 3rd Conf. on AGI.* http://agi-conf.org/2010/wp-content/uploads/2009/06/agi10singmodels2.pdf

[8] Vinge, V. (1993). 'The coming technological singularity.' *Whole Earth Review*, Winter issue.

[9] Singularity Institute for Artificial Intelligence. 'What is the singularity?' singinst.org/overview/whatisthesingularity . Accessed 20 Feb 2011.

[10] Singularity Institute for Artificial Intelligence. 'Why work toward the singularity?' singinst.org/overview/whyworktowardthesingularity. Accessed 20 Feb 2011.

[11] Brock, D.C. (ed) (2006). *Understanding Moore's Law: Four Decades of Innovation.* Philadelphia: Chemical Heritage Press.

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

[13] Metzinger, T. (2009), *The Ego-Tunnel: The Science of the Mind and the Myth of the Self* . NY: Basic Books.

[14] Holland, O. & Goodman, R (2003), 'Robots With Internal Models: A Route to Machine Consciousness?' *Journal of Consciousness Studies,* 10(4-5), 77-109

[15] Baars, B. J., & Franklin, S. (2009). 'Consciousness is computational: The LIDA model of Global Workspace Theory.' *Int. J. Machine Consciousness*, 1(1), 23-32.

[16] Ramamurthy, U and Franklin, S (2011). 'Self System in a Model of Cognition'. *Proc. MC 2011* (this Proceedings).

[17] Shanahan, M.P. (2005). 'Global Access, Embodiment, and the Conscious Subject', *Journal of Consciousness Studies*, 12(12), 46-66.

[18] Shanahan, M.P. (2010), *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*, Oxford: Oxford University Press

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

[20] Franklin, S. (2003). 'IDA: A Conscious Artefact?' *Journal of Consciousness Studies*, 10(4-5), 47-66.

[21] Torrance, S.B. (2007). 'Two conceptions of machine phenomenality', *Journal of Consciousness Studies*, 14(7), 154-166.

[22] McGurk, H. & McDonald, J. (1978). 'Hearing lips and seeing voices.' *Nature*, 264, 746-748.

[23] Tononi, G. (2008) 'Consciousness as integrated information: A provisional manifesto.' *Biological Bulletin,* 215: 216-242.

[24] Goertzel, B & Wang, P (eds.) (2007) *Advances in Artificial General Intelligence*, Amsterdam: IOS Press.

[25] Franklin, S (2007) 'A foundational architecture for artificial general intelligence.' In [24], 349-379.

[26] Goertzel, B. (2010). 'Toward a Formal Characterization of Real-World General Intelligence'. *Proc. 3rd Conf. Art. Gen. Intelligence*. . http://agi-conf.org/2010/wp-content/uploads/2009/06/paper_14.pdf

[27] Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences* N.Y.: Basic Books.

[28] Koch, C., and G. Tononi. 2008. 'Can machines be conscious?' *IEEE Spectrum* 45: 55–59.

[29] Seth A K, Izhikevich E, Reeke G N, Edelman G M. (2006). 'Theories and measures of consciousness: an extended framework.' *PNAS* 103(28): 10799—804

[30] Arrabales, R, Ledezma, A. and Sanchis, A. (2010). 'ConsScale: A pragmatic scale for measuring the level of consciousness in artificial agents.' *Journal of Consciousness Studies* 17(3–4), 131–64

[31] Raizer, K., Paraense, A. and Gudwin, R.(2011) 'A cognitive neuroscience-inspired codelet-based cognitive architecture for the control of artificial creatures with incremental levels of machine consciousness'. *Proc. MC 2011* (this Proceedings).

[32] Sloman, A. (1985) 'What enables a machine to understand?' *Proc Int. Joint Conf. A.I.* , Los Angeles, CA, 995-1001.

[33] Torrance, S. (1986). 'Ethics, mind and artifice' in K.S.Gill (ed) *Artificial Intelligence for Society*, Chichester: John Wiley, 55-72.

[34] Torrance,S. (2008) 'Ethics and consciousness in artificial agents', *Artificial Intelligence and Society*, 22(4), 495-521.

[35] Levy, D. (2009). 'The Ethical Treatment of Artificially Conscious Robots'. *Int. Jnl. Social Robotics*. 1, 209-216

[36] Wallach, W, Allen, C. And Franklin, S. (forthcoming). 'Consciousness and ethics: Artificially conscious moral agents'. *Int. Jnl. Machine Intelligence.* In press.

[37] Yukdowsky, E. and Singularity Institute for Artificial Intelligence (2001): *Creating Friendly AI 1.0.* http://singinst.org/ourresearch/publications/CFAI/index.html Accessed 20 Feb 2011

[38] Hibbard, B. (2001). 'Super-intelligent machines'. *ACM SIGGRAPH Computer Graphics*, **35**(1).

[39] Hibbard, B. (2004). 'Reinforcement learning as a Context for Integrating AI Research.' *2004 AAAI Fall Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research.*

[40] Coeckelbergh, M. (2009). 'Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents.' *Artificial Intelligence & Society*, 24, 181-189.

[41] Coeckelbergh, M (2010). 'Robot Rights? Towards a Social-Relational Justification of Moral Consideration', *Ethics and Information Technology* 12(3), 209-221

Proceedings of AISB '11: Machine Consciousness
Dimitar Kazakov and George Tsoulas (eds.)
ISBN 978-1-908187-06-2

Published by the Society for the Study of Artificial Intelligence and the Simulation of Behaviour
Printed by the University of York, York, UK