Biologically Inspired Cognitive Architectures II:
Papers from the AAAI Fall Symposium (FS-09-01)

Taking a Mental Stance Towards Artificial Systems

David Gamez and Igor Aleksander

Department of Electrical Engineering, Imperial College, London SW7 2BT, UK
dgamez@imperial.ac.uk, i.aleksander@imperial.ac.uk

Abstract

This paper argues that supervised cognitive growth in artifacts will be very difficult to achieve without detailed knowledge about systems' internal states. Physical information is too low level to provide a useful understanding of a system's behavior, and it is more pragmatically useful to take a mental stance towards an artificial system and interpret its actions in terms of mental states. This mental stance is similar to Dennett's intentional stance, except the ascription of beliefs and rationality in the intentional stance is replaced by the attribution of low level mental states in the mental stance. In some cases it might also be useful to take a conscious stance towards an artificial system that interprets its behavior as the outcome of a conscious decision making process. Since most artifacts lack language, automatic analysis techniques have to be used to identify the contents of their minds, and the second half of this paper suggests how some of the earlier work of Aleksander and Atlas can be applied in this area.

Introduction

A biologically-inspired cognitive system could be developed using a single black box approach. Only the external behavior would be monitored, and if the system failed to behave correctly, the parameters of the learning algorithm could be adjusted or the architecture changed, and the cycle repeated until the system behaved in the desired way. With the single black box approach it is difficult to identify the reasons for aberrant behavior, to know whether the system is overgeneralizing and to correct the architecture or learning parameters in appropriate ways. The single black box approach is also likely to become increasingly unworkable as the complexity of the system increases.

A more promising alternative is to develop cognitive systems using modular black boxes that are designed for different tasks and trained and tested independently. For example, if the system has a visual module and a motor module, then these could be trained and tested separately and then combined to produce the complete system. The system's developers could focus on the identification of faulty modules and only adjust and retrain the ones that did not work. A first limitation of the modular black box approach is that researchers are unlikely to know the expected behavior of every module in the system: if they did have this information, then there would be little need for a biologically inspired cognitive architecture that learnt from its experiences.[1] A second problem is that the training of the modules is likely to depend on the interactions between them, which would make it impossible to train and test them in isolation – in other words, modular black boxes may merge into a single black box. A third issue is that, as training progresses, the overall behavior of the system is likely to become a complete mystery, even if specific modules have been designed and there are known connections between them. Finally, biologically-inspired cognitive architectures might have to allow for the possibility that modules are used in several different ways depending on context, which would seriously complicate any attempt to completely specify their function.

These limitations of black box methodologies can be addressed by monitoring what is going on inside cognitive systems as they are being developed. This could be done by examining the low level interactions of the system – for example, in a neural system information about the neurons and their connections could be used to understand what the system has learnt and predict what it is going to do next. The problem with this type of low level approach, which is analogous to what Dennett (1987) calls the physical stance, is that it takes a great deal of effort and gives little insight into a system's behavior. A more promising approach is to interpret a system's actions as the outcome of a rational decision process based on beliefs and desires. This perspective on the system is what Dennett (1987, p. 17) calls the intentional stance:[2]

> … there is yet another stance or strategy one can adopt: the intentional stance. Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place

[1] One reason for the failure of De Garis' (2002) CAM-Brain system was that it required a full specification of all of the modules' functions.

[2] Dennett's design stance has been left out of this discussion because in a learning system most of the behavior is the result of training.

> in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many – but not all – instances yield a decision about what the agent ought to do; that is what you predict the agent will do.

Whilst the intentional stance is a good way of understanding the behavior of a person or animal, the attribution of rationality to the half formed mind of an experimental system undergoing training is unlikely to give much insight or have much predictive utility. Instead, this paper puts forward a version of the intentional stance called the mental stance, which takes the limited nature of artificial cognitive systems into account. The mental stance interprets a system as something that has a mind, and then frames its knowledge about the system in terms of mental states, representational mental states, relationships between mental states and cognitive functions. Since our current cognitive systems are incapable of expressing their states in language, third-person analysis techniques have to be used to map out a system's mind. Taking the mental stance enables us to see inside a system's mind as it is being trained, which makes it much easier to debug its learnt representations and understand its behavior.

In some cases it might be worth moving beyond the mental stance and taking a conscious stance towards a system, which interprets its actions as the outcome of a conscious decision-making process. Different theories about consciousness will make different predictions about a system's conscious states and it is an empirical question about which theory of consciousness provides the best understanding of a system's behavior.

The mental, intentional and conscious stances are adopted on pragmatic grounds because of the predictive power that is gained by interpreting a system in a particular way. The utility of these stances is independent of any actual mental, intentional or conscious states that a system might or might not have - questions about whether the system *really* has a mind, intentions or consciousness are set aside when viewing the system from one of these perspectives. Dennett's own position is realistic to the extent that he believes that the patterns interpreted as intentional are objectively present in the world, but he acknowledges that these patterns are just one particular way of looking at reality:

> I claim that the intentional stance provides a vantage point for discerning similarly useful patterns. These patterns are objective – they are *there* to be detected – but from our point of view they are not *out there* entirely independent of us, since they are patterns composed partly of our own "subjective" reactions to what is out there; (Dennett, 1987, p. 39)

The first part of this paper sets out some of the ways in which the mental stance could help us to understand a system's mental states. The next two parts explain how earlier work by Aleksander and Atlas could be used to understand the minds of cognitive systems.

Aspects of an Artificial Mind

Mental States

The mental stance interprets a system as something that has a mind. Although many philosophers would argue that mental states are conceptually distinct from physical states, the increase in our knowledge about the brain, and the constant reduction of mental functions to brain functions has led Churchland (1989) to suggest that the term "mental state" will eventually become redundant and mental terminology will be superseded by descriptions in terms of states of the brain – a position known as eliminative materialism. In the human case, this may eventually occur because a clear link has been established between mental states and the brain. However, when we take the mental stance towards an artificial system it is far from clear which part of the system we should examine when we interpret it as a mental entity. Within this context we need the concept of a mental state to specify the part of the system (or subset of the system's states) that is linked to states of its mind.

When people analyze human minds they generally focus on the brain and human mental states are usually taken to be states of human brains. However, other parts of the human body can be treated as mental as well. For example, states of the liver or blood could be interpreted as mental states and used to develop an understanding of a human's mind.[3] In artificial systems a mental state could be the firing activity in simulated neurons or the 1s and 0s in a computer's RAM - for example, mental states could be monitored in Franklin's IDA (2003) by using a debugger to measure memory changes. Different ways of defining a system's mental states may lead to different predictions about its behavior, and by experimenting with different mental state definitions we can identify the ones that provide the most accurate and efficient understanding of a system.

Representational Mental States

Systems that interact with their environment are likely to have representational mental states that co-vary with states of the world. In artificial systems these representations are likely to be very different from human beliefs and many of them will be difficult or impossible to express in natural language. Some systems will also have mental states that co-vary with states of their bodies, which can be interpreted as emotions using Damasio's (1995) theories.

A great deal of work has been carried out on the identification of representational mental states in the brain.

[3] See Paton et al. (2003) for a discussion of the computations carried out by the liver and other tissues.

Classic work in this area was carried out by Hubel and Wiesel (1959), who inserted electrodes into the brains of cats and measured the neural activity when different stimuli were present in different parts of the visual field. Neurons whose activity changed when an external stimulus was presented were judged to be representing the information in the stimulus. More recently fMRI scanning has been used to identify representational mental states - for example, Kay et al. (2008) used fMRI data to predict novel images that subjects were viewing with over 70% accuracy.

Less work has been carried out on the identification of representational mental states in artificial systems, although a number of techniques have been developed. For example, Krichmar et al. (2005) used a backtracing method to identify a system's functional pathways and Gamez (2008) used a combination of noise injection and mutual information to map out a network's representational states.

Relationships Between Mental States

The relationships between a system's mental states can significantly affect its perception. For example, a system that integrates color and shape information can represent a red cube. However, if the system cannot integrate color and shape information together, then it cannot represent the fact that a cube is red or imagine the possibility that a red cube could take on a different color. The relationships between mental states also determine the functions that are available in a system.

The relationships between mental states can be identified using methods for measuring functional and effective connectivity, such as Granger causality (Seth and Edelman, 2007), neural complexity (Tononi, Sporns and Edelman, 1994) or information integration (Tononi and Sporns, 2003; Balduzzi and Tononi, 2008). We are currently developing a new approach based on liveliness, which is covered later in this paper.

Cognitive Functions

Knowledge about the presence or absence of cognitive functions, such as planning, imagination and attention, is essential if we want to understand a system's training and behavior. Whilst some work has been carried out on validating functions using external behavior – unit testing in software development, for example – cognitive functions are more likely to be identified by examining a system's internal states. One approach would be to define a function in terms of a mapping between input and output and look for groups of mental states that implement this mapping. The problem with this method is that all possible combinations of input would have to be included in the definitions, which would be impossible for real valued inputs. A more promising approach would be to define functions in terms of workflows and use the relationships between mental states over time to determine whether a particular function is possible or active in the system. Some suggestions about how this approach could work are given towards the end of this paper.

Conscious Mental States

In some cases it might be worth going beyond the mental and intentional stances and take a conscious stance towards a system, which interprets its actions as the outcome of a conscious decision making process. In many cases a conscious stance might be able to provide a more accurate and efficient way of understanding a cognitive system's training and make better predictions about its behavior. Since most contemporary artificial systems are unable to report their conscious states directly, an analysis process would have to be used to predict their conscious states. Different theories of consciousness are likely to make different predictions,[4] and it is a pragmatic question about which theory would give the most useful understanding of a system's learning and behavior.[5]

Descriptions of Mental States

The mental stance enables a system's knowledge and behavior to be concisely summarized using high level descriptions of its mental states. However, Nagel (1974) and Chrisley (1995) have raised a number of problems with natural language descriptions of non-human and artificial mental states, which suggest that new methods of description may have to be found.

Previous work in this area includes graphical representations of the inner states of a robot (Holland and Goodman, 2003; Stening et al., 2005), and Chrisley and Parthemore's (2007) use of a SEER-3 robot to specify non-conceptual mental content. Ascoli and Samsonovich (2008) have developed an approach to this problem using semantic maps and a companion paper (Aleksander and Gamez, 2009) demonstrates how iconic representations can be used to dynamically display a system's inner states.

Liveliness in Neural Networks

Weightless and Spiking Neurons

The earlier work of Aleksander (1973) and Aleksander and Atlas (1973) on a logic-based approach to neural networks is used in this half of the paper to suggest some ways of analyzing an artificial mind. This approach is applicable to neurons whose function can be expressed as a truth table – for example, weightless neurons (Aleksander et al., 2009) or simple models of spiking neurons. To see how spiking neurons can be expressed as truth tables, consider a simple neuron with the following features:

- Neuron can emit 1 (spike) or 0 (no spike).

[4] Early work in this area can be found in Gamez (2008).

[5] The conscious stance is unlikely to be able to explain all of a system's behaviors because many human actions take place automatically without conscious control, and the same is likely to be true of artificial systems.

- When the neuron receives a spike on one of its connections, the weight of that connection is added to the neuron's membrane potential.
- If the value of the membrane potential exceeds the threshold, the neuron fires and emits 1. Otherwise the neuron emits 0.
- Spikes from previous time steps are not taken into account – in effect the time step is set to a large value so that the membrane potential is reset after each time step.

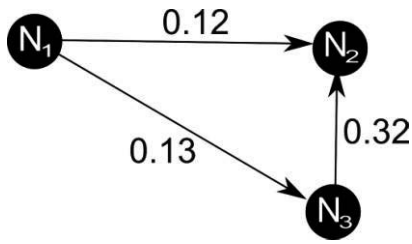

Figure 1: Example spiking network

In the example network shown in Figure 1, information about the weights enables the function of N_2 to be expressed in the truth table shown in Figure 2.

N_1	N_3	N_2 (0.1)	N_2 (0.3)	N_2 (0.5)
0	0	0	0	0
0	1	1	1	0
1	0	1	0	0
1	1	1	1	1

Figure 2: Truth table for N_2 in Figure 1. The output of N_2 depends on the threshold, which is given in brackets

A truth table representation of a neurons' functions enables a number of analytical results to be derived, which are covered next.

Liveliness

Liveliness is a key concept for the work described in this paper because it measures the probability that a connection or neuron transmits information. With a lively connection there is a high probability that a change of signal in the input will result in a change in the output, irrespective of the inputs to the other connections. On the other hand, the input to a non-lively connection will have little or no effect on the output of the neuron. For example, in the network shown in Figure 1, the connection between N_3 and N_2 has a high liveliness for thresholds of 0.1 and 0.3 because a change in N_3 is almost always reflected in a change in the output of N_2. On the other hand, the connection from N_1 to N_2 has low liveliness because a change in N_1 is rarely reflected in a change in the output of N_2, and only when the threshold is 0.1.

The probability, P_j, that a connection transmits a signal can be calculated by dividing the number, m, of conditions under which it transmits a signal by the number of possible states of the other K inputs to the neuron, as shown in Equation 1:

$$P_j = \frac{m}{2^{K-1}} \tag{1}$$

The liveliness of a neuron, λ_k, is the average probability of it being live (in a signal transmission sense) between an arbitrarily selected input and its output, with the average being taken over the K connections to the neuron, as expressed in Equation 2:

$$\lambda_K = \frac{1}{K}\sum_j P_j \tag{2}$$

It is also possible to work out the liveliness of a set of neurons whose functions are drawn with equal probability from some set F:

$$L_F = \frac{1}{F}\sum_f \lambda_K \tag{3}$$

In Equation 3 L_F is the average probability of a neuron being live between an arbitrarily selected input and its output given that the neuron's function is selected at random from set F.

Cyclical activity in a network only occurs within rings of lively elements that have a high probability of transmitting signals to each other. Using the results given above, Aleksander (1973) derives a formula expressing the number, R_l^N, of live rings with l elements in a network of N neurons:[6]

$$R_l^N = \frac{N!}{l!(N-l)!} * \left(\frac{K}{N} - \frac{1}{N^2} - \frac{1}{N^3} \cdots - \frac{1}{N^K}\right) * (L_F)^l \tag{4}$$

Equation 4 can be used to find the distribution of the lengths of lively rings. On the assumption that the lively rings are independent, the total number of network states involved in cyclic activity is the product of the number of possible states in each ring, as shown in Equation 5:

$$S_{tot} = \prod_{j=1}^{n} 2^{l_j}, \tag{5}$$

where n is the number of isolated live rings and l_j is the length of ring j. The detailed derivation of these results can be found in Aleksander (1973) and Aleksander and Atlas (1973). The application of these results to the analysis of a system's mental states is covered next.

Applications of the Liveliness Approach

Representational Mental States

The liveliness of connections and elements provides a way of understanding the relationship between internal activity of the system and signals that reach the system from the

[6] Equation 4 is slightly different from Aleksander (1973) because it takes networks with a connectivity (K) greater than 2 into account.

outside world. For example, an analysis for representational mental states could be carried out using a modified version of Krichmar et al.'s (2005) backtracing method, which would start with an internal neuron whose representations were unknown and identify neurons with lively connections to this neuron. By repeating this process it would be possible to move back through the network until neurons were found with known representational states – perhaps because they were directly connected to sensory input or motor output.

Information Integration

The liveliness of a connection indicates how likely it is that information will be exchanged between two neurons and this can be extended through intermediate points to work out the information relationships between indirectly connected neurons. So, for example, questions about whether the system is aware of a red cube or separately aware of redness and cubeness can be answered by looking at the liveliness of the paths connecting the representations of redness and cubeness. It would also be possible to use the distribution of lively rings in different parts of the network to measure the differentiation of a system's states, and the liveliness approach might open up a way of calculating the Φ measure of information integration that avoids the processor-heavy factorial dependencies of Tononi and Sporns' (2003) and Balduzzi and Tononi's (2008) methods.

Learning

The distribution of information integration in a network could be a useful way of monitoring its learning. In a typical training scenario a system starts off with a high level of connectivity between its elements. As learning proceeds the strength of these connections is adjusted so that the connection patterns reflect the co-occurrence of events in the world. The untrained system will have high integration and low differentiation, which will result in a low value of Φ. As training proceeds, the integration will decrease as weights are reduced on certain connections, but the differentiation will increase because only particular combinations of states will occur - resulting in a net increase in the system's Φ. However, if training goes on for too long, then the integration is likely to become too low and the value of Φ will go down – suggesting that there is a 'sweet spot' in which the balance between differentiation and integration is maximized. It would also be possible to use the number of lively rings and the richness of the state space to measure a system's training.[7]

Cognitive Functions

The liveliness measures could be used to identify rings and chains of mental states that correspond to different functions: a lively ring of mental states would be a repeating function; a lively chain of mental states would be a linear function. This approach to function definition would make it possible to know whether, for example, the system is using its imagination to solve a particular problem, and one could map out the possible functions of a system and decide whether it has learnt a particular function.

Conscious Mental States

The liveliness approach offers a number of ways of making predictions about conscious mental states according to different theories of consciousness. To begin with, Tononi (2008) makes an explicit connection between Φ and consciousness: if the liveliness approach could be used to calculate Φ, then it could also be used to make predictions about which mental states are associated with consciousness. Secondly, Metzinger (2003) links consciousness with information integration over time, which he calls the window of presence (Constraint 2) and dynamicity (Constraint 5). Since lively rings store earlier states of the system, the distribution of lively ring lengths could be used to measure the system's window of presence and dynamicity. The state-based approach to functions also makes it possible to move from high level theories about consciousness to low level information about the system, allowing predictions to be made about whether the system conforms to Aleksander's (2005) five axioms of depiction, imagination, volition, attention and emotion.

Future Work

This work linking liveliness and a truth table-based analysis to the mental and conscious stances is at a very early stage of development and a great deal more research is required. At the moment our primary focus is on the development of an open source software package that can carry out the liveliness and Φ calculations. We are planning to use this software to analyze the liveliness and integration of simulated networks and this will be compared with the results of Balduzzi and Tononi (2008). Some preliminary experiments in this area can be found in a companion paper (Aleksander and Gamez, 2009).

Conclusions

This paper has argued that single or modular black box approaches will be inefficient and laborious ways of developing the next generation of cognitive systems. Taking the mental stance and viewing a system as if it has a mind provides a much better understanding of a cognitive system's learning and behavior, and in some cases it might be useful to take a conscious stance towards a system.

The first half of this paper outlined how the mental stance could be based on an analytical identification of mental states, representational mental states, relationships between mental states and cognitive functions. The second half made some suggestions about how the work of

[7] The relationship between information integration and learning is very speculative at this stage and more empirical work is needed.

Aleksander (1973) and Aleksander and Atlas (1973) could be used to characterize a system's mental contents. The mental stance and the conscious stance are pragmatic positions and their utility is entirely separate from questions about whether the system *really* has a mind or conscious mental states.

Acknowledgements

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

References

Aleksander, I. 1973. Random Logic Nets: Stability and Adaptation. *International Journal of Man-Machine Studies* 5: 115-31.

Aleksander, I. 2005. *The World in My Mind, My Mind in the World: Key Mechanisms of Consciousness in People, Animals and Machines*. Exeter: Imprint Academic.

Aleksander, I. and Atlas, P. 1973. Cyclic Activity in Nature: Causes of Stability. *International Journal of Neuroscience* 6: 45-50.

Aleksander, I., França, F., Lima, P., and Morton, H. 2009. A brief introduction to Weightless Neural Systems. *Proceedings of ESANN 2009*, Bruges.

Aleksander, I. and Gamez, D. 2009. Iconic Training and Effective Information: Evaluating Meaning in Discrete Neural Networks. *Proceedings of the AAAI Fall Symposium on Biologically Inspired Cognitive Architectures.*

Ascoli, G. A. and Samsonovich, A.V. 2008. Science of the Conscious Mind. *The Biological Bulletin* 215: 204-215.

Balduzzi, D. and Tononi, G. 2008. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput. Biol.* 4(6).

Chrisley, R. J. 1995. Taking Embodiment Seriously: Nonconceptual Content and Robotics. In K.M. Ford, C. Glymour and P.J. Hayes (eds.), *Android Epistemology*. Menlo Park, Cambridge and London: AAAI Press/ The MIT Press.

Chrisley, R. J. and Parthemore, P. 2007. Synthetic Phenomenology: Exploiting Embodiment to Specify the Non-Conceptual Content of Visual Experience. *Journal of Consciousness Studies* 14 (7): 44-58.

Churchland, P. 1989. *A Neurocomputational Perspective*. Cambridge, Massachusetts: The MIT Press.

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

De Garis, H. and Korkin, M. 2002. THE CAM-BRAIN MACHINE (CBM) An FPGA Based Hardware Tool which Evolves a 1000 Neuron Net Circuit Module in Seconds and Updates a 75 Million Neuron Artificial Brain for Real Time Robot Control. *Neurocomputing* 42(1-4): 35-68.

Dennett, D. C. 1987. *The Intentional Stance*. Cambridge, Massachusetts: MIT Press.

Franklin, S. 2003. IDA: A Conscious Artifact. In O. Holland (ed.), *Machine Consciousness*. Exeter: Imprint Academic.

Gamez, D. 2008. *The Development and Analysis of Conscious Machines*. Unpublished PhD thesis, University of Essex, UK. Available at: www.davidgamez.eu/mc-thesis.

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

Hubel, D. H. and Wiesel, T.N. 1959. Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology* 148(3): 574–91.

Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. 2008. Identifying natural images from human brain activity. *Nature* 452, 352–355.

Krichmar, J. L., Nitz, D. A., Gally, J. A. and Edelman, G. M. 2005. Characterizing functional hippocampal pathways in a brain-based device as it solves a spatial memory task. *PNAS* 102(6): 2111-6.

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

Nagel, T. 1974. What is it like to be a bat? *The Philosophical Review* 83: 435-56.

Paton, R., Bolouri, H., Holcombe, W. M. L., Parish, J. H. and Tateson, R. 2003. *Computation in Cells and Tissues: Perspectives and Tools of Thought*. Berlin and Heidelberg: Springer-Verlag.

Seth, A. K. and Edelman, G. M. 2007. Distinguishing causal interactions in neural populations. *Neural Computation* 19(4): 910-33.

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

Tononi, G. 2008. Consciousness as Integrated Information: a Provisional Manifesto. *Biological Bulletin* 215: 216-242.

Tononi, G. and Sporns, O. 2003. Measuring information integration. *BMC Neuroscience* 4:31.

Tononi, G., Sporns, O. and Edelman, G. M. 1994. A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. USA* 91: 5033-7.