International Journal of Machine Consciousness


# THE INFORMATIONAL MIND AND THE INFORMATION INTEGRATION THEORY OF CONSCIOUSNESS

DAVID GAMEZ

*Department of Informatics, University of Sussex,*
*Brighton, BN1 9RH, UK*
*david@davidgamez.eu*

This is the author-generated version of an article published in the *International Journal of Machine Consciousness* 6(1): 1-8.. The original publication is available at:
http://www.worldscientific.com/worldscinet/ijmc.

**Abstract**. According to Aleksander and Morton's informational mind hypothesis, conscious minds are state structures that are created through iconic learning. Distributed representations of colors, edges, objects, etc. are linked with proprioceptive and motor information to generate the awareness of an out-there world. The uniqueness and indivisibility of these iconically learnt states reflect the uniqueness and indivisibility of the world. This article summarizes the key claims of the informational mind hypothesis and considers them in relation to Tononi's information integration theory of consciousness. Some suggestions are made about how the informational mind hypothesis could be experimentally tested, and its significance for work on machine consciousness is considered.

*Keywords*: Aristotle's Laptop; information integration; informational mind; correlates of consciousness.

## 1. Introduction

*Aristotle's Laptop* sets out Aleksander and Morton's [2012] informational mind hypothesis about what it is to be a conscious thinking being. This draws on Aleksander's earlier axiomatic work [Aleksander, 2005] as well as Tononi's information integration theory of consciousness [Tononi, 2004; 2008], and it is applicable to both humans and machines.

The book is highly readable and contains useful background material and short biographical sketches of the key figures. Neural models are often used to illustrate the main ideas, which function as an alternative and (for some) more intuitive way of describing the phenomenon. While the authors are careful to stress that their models are not exact simulations of the brain, they are a good way of pointing to aspects of brain function that might instantiate the mechanisms they discuss. This halfway house between theoretical ideas about consciousness and the brain could be a useful first step towards testing the informational mind hypothesis in the future.

The first part of this review summarizes the informational mind hypothesis, which is contrasted with Tononi's information integration theory of consciousness in Section 3. I will then explore how it could be experimentally tested and finish with a brief discussion of its implications for machine consciousness.

## 2. The Informational Mind Hypothesis

The informational mind hypothesis summarizes the approach to the mind and consciousness that Aleksander and Morton develop in *Aristotle's Laptop*. It is stated as follows:

> The mind of an entity is the *state structure* of specialized parts of a neural brain (living or artificial), created through *iconic learning*. These areas are characterized by two features. First, they are able to create world-representing states by compensating for the muscular action of the body which is used in exploring this world. Second, such specific areas of the brain have interconnections that exhibit sufficiently high values of *information integration* to ensure both that the states of the state structure generate information by being *unique* and that they are capable of representing causal relationships by being *indivisible.* [Aleksander & Morton, 2012, p. 152-153]

This section briefly summarizes the main elements of the informational mind, which are discussed in relation to Tononi's information integration theory of consciousness in Section 3.[*]

### 2.1. *Iconic Learning and State Structure*

Iconic learning is a way in which a neural network is trained to respond to patterns in the world. At each point in time the network is exposed to a set of input data and some or all of this data is copied into the neurons' states [Aleksander, 1996]. This type of learning is a kind of picturing of the external world, which Aleksander and Morton contrast with more abstract representations in which, for example, a single neuron encodes an object or image. However, iconic learning does not imply that a neat picture is formed in a neural network – it is only necessary for the data to be mapped onto the neurons in some arbitrary but consistent way.

In Aleksander and Morton's examples of iconic learning the neurons are trained with a substantial amount of feedback from other neurons, which makes the learnt states stable attractors. This enables the network to reconstruct patterns from noisy or incomplete examples, and the feedback also enables the network to learn sequences of patterns. If the network is disconnected from its input after training, it can enter sequences of states that are similar to dreaming or imagination.

Aleksander and Morton claim that the spatiotemporal structure which the network acquires through iconic learning closely corresponds to our phenomenological experience, and it enables the network to mirror the spatiotemporal structure and causal relationships of the world. This provides a flexible way in which the organism can reason about the world and make plans, which is not facilitated by more abstract and specialized representations.

[*] I apologize for any mistakes in this summary of Aleksander and Morton's theories.

### 2.2. *Depiction*

As Aleksander and Morton discuss in Chapter 7, the simple examples of iconic learning that are demonstrated in some of their models are substantially different from the learning and representation in the brain. While the book's models are typically based around a layer of 98 x 98 neurons trained with black and white images, the human visual system consists of many separate topographic maps encoding different types of information, which are distributed across different brain areas. This presents the problem about how the separate representations in the brain are connected together into a single phenomenological experience. Drawing on their earlier work on depiction [Aleksander, 2005], Aleksander and Morton suggest that proprioceptive and motor information is the binding element that links information from different brain areas into an experience of an out-there world. For example, proprioceptive and motor information is used to bind information about color, movement and objects in separate brain areas to a single region of physical space.

### 2.3. *Uniqueness and Indivisibility*

Aleksander and Morton's concepts of uniqueness and indivisibility are based on Tononi's work on information integration, which is summarized in Chapter 6. Uniqueness is the amount of information that is associated with a system being in a particular state, which varies with the number of possible states of the system. For example, a state of a system with a single binary element has a low amount of uniqueness because each state is only differentiated from one other state. On the other hand, each state of a system with four binary elements is much more unique because it is distinguished from the other 15 states that the system could be in. In terms of information theory (see Chapter 2), the information that is gained by knowing the state of a four element system is 4 bits, whereas the information that is gained by knowing the state of a single element system is only 1 bit. According to my understanding of what Aleksander and Morton mean by uniqueness, the probability distribution of the states should also affect the uniqueness of each state. A system is more likely to be in states with higher than average probability, and so knowing that a system is in a high probability state is less informative than knowing that it is in a low probability state. All of this suggests that uniqueness varies with the number of neurons in a network, the number of states of each neuron and the probability distribution of the network's states.

Indivisibility corresponds to what Tononi calls integration, which is the extent to which the individual elements of a system causally interact to produce a particular state. This corresponds to the first-person observation that our conscious experiences are indivisible wholes in which everything is bound together. In the informational mind indivisibility is linked to the causal interactions between the elements that result from iconic learning.

An untrained weightless neural network randomly shifts between states. It has maximum uniqueness because all of the states are equally probable, but there is no

indivisibility because there is no integration or causal relationships between the states. As the network is iconically trained some of the states become stable attractors, which increases their probability and reduces their uniqueness.[†] At the same time, the learnt causal relationships between the states will increase the indivisibility of the system. It is not clear where the balance between uniqueness and indivisibility reaches its peak in a given system, which is why Tononi has proposed algorithms that measure this balance and identify the parts of the system where it is maximized [Tononi & Sporns, 2003; Balduzzi & Tononi, 2008].

## 3. Contrast between the Informational Mind Hypothesis and the Information Integration Theory of Consciousness

As Aleksander and Morton explain in Chapter 6, Tononi [2004; 2008] has developed a theory about the relationship between information and consciousness that has had a significant influence on *Aristotle's Laptop*. The information integration algorithms put forward by Tononi [2003; 2008] can identify the part of a system that maximizes the balance between uniqueness and indivisibility, which Tononi claims is the area associated with consciousness. These algorithms output a number, Φ, that is predicted to correspond to the amount of consciousness in the system. Tononi has also proposed a method for generating a high dimensional structure that is hypothesized to correspond to the contents of consciousness [Tononi, 2008]. A key limitation of Tononi's information integration algorithms is that they have factorial dependencies, which makes it impractical to apply them to systems with more than ~20 elements. To address this problem, other ways of calculating information integration have been put forward, including the liveliness algorithm developed by Gamez and Aleksander [2011].

One significant issue with Tononi's theory is his claim that consciousness is identical with information integration. Since stones have internal states that exhibit some degree of information integration, this effectively asserts that consciousness can be found everywhere. The deeply sleeping brain will also have some amount of information integration, which contradicts our observation that we are unconscious in this state.[‡] Aleksander and Morton's richer account of the informational mind is better at avoiding panpsychism because it applies to a much more limited set of systems – for example, the states of a stone are unlikely to be depictive.

Another limitation of Tononi's approach is that it lacks a clear definition of information. Floridi [2009] makes a nice distinction between data and information as meaningful data, which suggests that Tononi's information integration theory of consciousness it is more accurately described as a theory of data integration [Gamez, 2011]. In Aleksander and Morton's account, the internal states of a system (for example, the states of the weightless neurons) are data in Floridi's sense. But since these states are related to each other and to the world, they can also plausibly be argued to be meaningful

[†] It also reduces the information entropy of the network.

[‡] This problem could be addressed by applying a threshold to Φ, but this would not be compatible with Tononi's claim that information integration *is* consciousness.

data, or information. This suggests that the informational mind is a hypothesis about the connection between information and consciousness, whereas Tononi's theory, which lacks an account of meaningful data, is not. This is not necessarily a limitation of Tononi's theory, unless you have independent grounds for believing that information and not data is important for consciousness.

There are also some problems with Tononi's geometric account of conscious contents, which fails to explain why our conscious experiences are so narrowly tied to our sensations, when in principle we could have completely novel and indescribable conscious experiences all the time. Furthermore, it is difficult to see how a geometric quale could be used to make predictions about the contents of consciousness, and it is not clear how radically the shape of a quale will be affected by minor changes in conscious contents. Aleksander and Morton's iconic representations provide a more plausible account of conscious contents that explains why consciousness contains world-driven information and why hallucinations consist of familiar sensations, such as sound and color.

Although Aleksander and Morton's link between consciousness and iconic learning addresses some of the limitations of Tononi's theory, it does introduce problems of its own. One of the more serious issues is that iconic states are only linked to the environment at the time of their creation: they can then be activated offline in dreams and imagination. So what is it about a learnt iconic state that makes it conscious when it is activated offline? Suppose that system A learns a sequence of iconic representations, and then a copy of system A is created and both systems are run offline in a completely different environment. The sequence of states in both systems will be identical, and so both will presumably be conscious to the same extent. This makes it hard to say why the learning matters, rather than just the state. This would support Tononi's view that consciousness is linked to the structural characteristics of the state, rather than to the fact that it is representational or has been learnt.

While philosophical arguments can be a useful method for comparing different theories, the best way to test a theory is to carry out experiments to check its predictions. Some preliminary experimental work has been carried out on the link between information integration and consciousness [Lee et al., 2009; Massimini et al., 2009; Ferrarelli et al., 2010]. Further research is needed to devise experiments that could test the predictions of the informational mind hypothesis, which will be discussed next.

## 4. Testing the Informational Mind Hypothesis

The standard procedure for experimentally testing a link between consciousness and the physical world is to measure consciousness, measure the physical world and look for correlations between the two [Gamez, 2012]. If the informational mind hypothesis is correct, it should be possible to show that an informational mind is active when a system is conscious, and inactive when a system is non-conscious or in the non-conscious parts of a system.

Before this type of experiment can be carried out, it is necessary to clarify the difference between a conscious informational mind and the unconscious parts of a system. In Chapter 9 Aleksander and Morton discuss the Freudian unconscious, which is described as a collection of mental states that have become inaccessible because of their highly charged negative emotional values. However, the authors choose not to discuss the other unconscious processes in the brain, such as unconscious driving or blindsight, which are more typically contrasted with conscious states in experiments on the neural correlates of consciousness. To clarify the distinction between these different types of unconscious I will use 'unconscious' to refer to the Freudian unconscious, and 'non-conscious' to refer to states of the brain that are not conscious in any way. When the brain is in deep sleep all of its states are non-conscious; in binocular rivalry experiments the input from one eye is conscious and the input from the other eye is non-conscious. Presumably unconscious states are for the most part non-conscious, except when they are recalled during therapy.

Experiments on the correlates of consciousness typically use a contrastive methodology that looks for differences in the conscious and non-conscious brain or between brain states that are processing conscious or non-conscious information [Tononi & Koch, 2008]. Since the mechanisms proposed by Aleksander and Morton are presumably only associated with conscious brains and information, it should be possible to use this type of contrastive experiment to test their theory. To do this it is necessary to specify in more detail what sorts of state machines are associated with non-conscious information. For example, the non-conscious states might be a separate state structure or a distinct system within the brain that is not reachable from the main conscious state structure. If the informational mind hypothesis is correct, the non-conscious states should lack some of the key features of the informational mind. While they will contain representations – for example, in the case of non-conscious driving these will include the road, steering wheel, etc. – these might not be iconic representations, and perhaps they will not be depictive and their structure will have less uniqueness and indivisibility than conscious states.

The informational mind hypothesis cannot be systematically tested if its presence or absence is measured using the subjective judgment of the experimenter. For example, a state cannot be judged to be depictive just because it is linked to a neural layer that is arbitrarily labeled 'motor data'. This problem can be addressed by expressing the informational mind hypothesis in a mathematical or algorithmic form. For example, graph theory or category theory could be used to express the degree to which a representation is iconic, and I have suggested elsewhere how depiction can be algorithmically measured [Gamez, 2008]. A mathematical version of the informational mind hypothesis could be tested in a similar way to Tononi's information integration approach.

## 5. Machine Consciousness and the Informational Mind

If the informational mind can be shown to be linked to consciousness in humans, then it might be claimed that *any* system that instantiates these mechanisms will be conscious, including artificial systems. Before this claim can be made, it is necessary to show that *any* implementation of the informational mind is linked to consciousness, and not just the implementation of the informational mind in the brain. For example, the informational mind might be necessary but not sufficient for consciousness - biological neurons might be necessary as well. In this case an implementation of an informational mind in a computer would not be conscious.

It is likely to be very difficult to devise experiments that could show that the informational mind is correlated with consciousness independently of the neural substrate. This type of claim might also run up against the objection that it leads to an untenable panpsychism [Bishop, 2002; 2009]. In my own work I have argued that anything that fits within the limits of what we can show to be correlated with consciousness in the human brain should be attributed consciousness [Gamez, 2012]. This would lead to consciousness being attributed to many artificial implementations of the informational mind.

## Acknowledgements

The writing of this review was supported by a grant from the John Templeton Foundation (ID 15619: 'Mind, Mechanism and Mathematics: Turing Centenary Research Project'). I would also like to thank the Sackler Centre for Consciousness Science at the University of Sussex for hosting me as a Research Fellow during this project.

## References

Aleksander, I. and Morton, H. [2012] *Aristotle's Laptop: The Discovery of Our Informational Mind* (World Scientific, Singapore).

Aleksander, I. [2005] *The World in My Mind, My Mind in the World* (Imprint Academic, Exeter).

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

Tononi, G. [2008] Consciousness as Integrated Information: A Provisional Manifesto, *Biol Bull* **215**(3), 216-242.

Aleksander, I. [1996] Iconic Learning in Networks of Logical Neurons, in *Evolvable Systems: From Biology to Hardware*, edited by T. Higuchi, M. Iwata and W. Liu (Springer, Berlin Heidelberg). 1259.

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

Balduzzi, D. and Tononi, G. [2008] Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework, *PLoS Comput Biol* **4**(6), e1000091.

Gamez, D. and Aleksander, I. [2011] Accuracy and Performance of the State-Based Φ and Liveliness Measures of Information Integration, *Consciousness and Cognition* **20**(4), 1403-1424.

Floridi, L. [2009] Philosophical Conceptions of Information, *Lecture Notes in Computer Science* **5363**, 13-53.

Gamez, D. [2011] Information and Consciousness, *Etica & Politica / Ethics & Politics*, **XIII**(2), 215-234.

Lee, U., Mashour, G. A., Kim, S., Noh, G. J. and Choi, B. M. [2009] Propofol Induction Reduces the Capacity for Neural Information Integration: Implications for the Mechanism of Consciousness and General Anesthesia, *Consciousness and Cognition* **18**(1), 56-64.

Massimini, M., Boly, M., Casali, A., Rosanova, M. and Tononi, G. [2009] A Perturbational Approach for Evaluating the Brain's Capacity for Consciousness, *Prog Brain Res* **177**, 201-214.

Ferrarelli, F., Massimini, M., Sarasso, S., Casali, A., Riedner, B. A., Angelini, G., Tononi, G. and Pearce, R. A. [2010] Breakdown in Cortical Effective Connectivity During Midazolam-Induced Loss of Consciousness, *Proc Natl Acad Sci U S A* **107**(6), 2681-2686.

Gamez, D. [2012] Empirically Grounded Claims About Consciousness in Computers, *International Journal of Machine Consciousness* **4**(2), 421-438.

Tononi, G. and Koch, C. [2008] The Neural Correlates of Consciousness: An Update, *Ann N Y Acad Sci* **1124**, 239-261.

Gamez, D. [2008]. *The Development and Analysis of Conscious Machines*. Unpublished PhD Thesis, University of Essex.

Bishop, J. M. [2002] Counterfactuals Cannot Count: A Rejoinder to David Chalmers, *Consciousness and Cognition* **11**(4), 642-652.

Bishop, J. M. [2009] A Cognitive Computation Fallacy? Cognition, Computations and Panpsychism, *Cognitive Computation* **1**, 221-233.