International Journal of Machine Consciousness

WORLD-RELATED INTEGRATED INFORMATION: ENACTIVIST AND PHENOMENAL PERSPECTIVES

MIKE BEATON

Dept. of Informatics, University of Sussex,
Brighton BN1 9QJ
mjsbeaton@gmail.com

IGOR ALEKSANDER

Dept. of Electrical and Electronic Engineering, Imperial College,
London SW7 3BT
i.aleksander@imperial.ac.uk

Received Day Month Year
Revised Day Month Year

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. 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. 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. 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). To have (or gain) information is to behave in a Bayesian-rational way in response to evidence. Information only ever belongs to whole, rationally behaving agents; information is only 'in the brain' from the point of view of a theorist seeking to explain behaviour. Rational behaviour (hence information) will depend on brain, body and world – embodiment matters. Then, 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, 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 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. 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.

Keywords: Integrated Information, Bayesian Rationality, Iconic Learning, Depiction.

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 [Shear and Jevning 1999]. 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 [Metzinger 2003].

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.

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 [Tononi and Sporns 2003; Balduzzi and Tononi 2008]. 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[a]. 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" [Tononi and Sporns 2003]

[a] 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.

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[b]. 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[c], 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 (normalized[d]) 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 certainly is a measure of information – this follows directly from

[b] 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.

[c] 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'.

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

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 [Balduzzi and Tononi 2008] 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)[e].

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[f]. 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.

[e] 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.

[f] 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.

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[g]. 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[h], 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) [Balduzzi and

[g] 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 [2011] 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.

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

Tononi 2008]. Elsewhere, he goes as far as to claim that Φ "exists as a fundamental quantity – as fundamental as mass, charge, or energy" [Tononi 2008].

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 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." [Balduzzi and Tononi 2008]

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 [Jaynes 2003] following Cox [Cox 1946] (and earlier writers, including Keynes [Keynes 1921]) 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) [Erickson and Smith 1988].

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 [Jaynes 2003] (following Shannon [Shannon 1948]) 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." [Jaynes 2003]

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

[i] 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.

defined as comparisons between probability distributions (the most simple information measure being just the arithmetical difference between 'before' and 'after' entropy values[j]).

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[k].

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 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 [Keynes 1921; Jaynes 2003]: 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.).

[j] 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.

[k] 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).

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[l], and then acting rationally on their subjective credences[m], 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 [2003] and Izquierdo and Di Paolo [2005] 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 [Izquierdo and Di Paolo 2005], 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" [Clark and Toribio 1994]; 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).

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

[m] 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) [Dennett 1987] and range of applicability.

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 [Aleksander and Morton 2007]. This has previously been discussed as 'an axiomatic theory of consciousness' [Aleksander 2007] 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[n]. 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 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:

[n] We argue for this usage of the word *phenomenal* in Aleksander and Morton [2007].

$I = \{i_1, i_2, ...i_{/I/}\}$, 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 [Aleksander, França et al. 2009] 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 $i_t, q_{t-1} \rightarrow q_t (= i_t)$. 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, generalization takes place in the sense that a pair $(i_a, q_a) \rightarrow i_j$ where $(i_j, q_j) \rightarrow i_j$ 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.

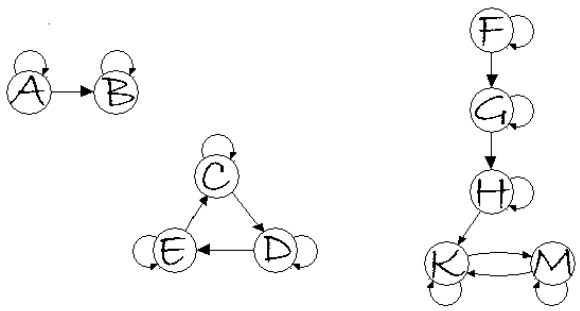

Figure 1: Structures of world states. The letters are 40x40 bit images at *I* and are not symbols. All states can 'linger' as well as transit to another state. So, looking at the top left behaviour, the world can sustain A from which it can change to B and rest there, but not return to A. Four behaviours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

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

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

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 [Aleksander and Gamez 2009], but disrupt the ability of the net to identify the state structure of the world.

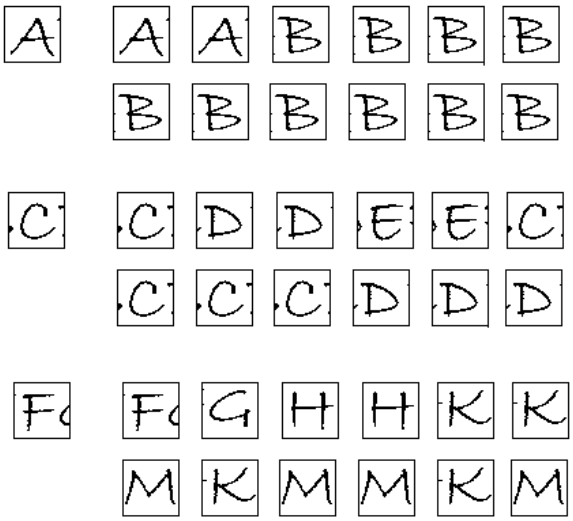

Figure 2: The inner response of a fully connected net. The left image is at the input for one time step only, after which it is replaced by noise to disconnect the automaton from the world. This noise then randomises the transition back into the current state or on to the next state.

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.[o] So it is seen that exposure to the state structure of the world is 'understood' by the learning network which identifies the four distinct behaviors of the world and generates internal representations that fully represent these behaviors. 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

[o] As an aside this has to be distinguished from a 'fully connected Hopfield network' as discussed by Balduzzi and Tononi [2008]. 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.

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.

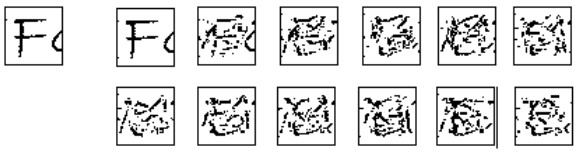

Figure 3. Now each neuron samples the output of 25 neurons (5 x 5 from the state array) and from a corresponding 5x5 array from *I*. The response to F in the third state group of fig.1 is shown and should be contrasted with the third result in fig.2. Each neuron samples the output of 25 neurons (5 x 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 [Balduzzi and Tononi 2009] 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.

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 behaviorally). We have suggested that these two levels may not just be logically distinct, but actually distinct, in the case of brains and behavior: 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 [Seth 2009] of something more fundamentally associated with consciousness (for instance, of the instantiation of consciousness understood axiomatically [Aleksander and Dunmall 2003], or simply of the presence of coherent, complex rational behavior [Beaton 2009]).

Acknowledgements

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

References

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

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

Aleksander, I., F. França, et al. [2009] *A brief introduction to Weightless Neural Systems*. Proceedings of ESANN 2009, Bruges.

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

Aleksander, I. and H. Morton [2007] Depictive Architectures for Synthetic Phenomenology. *Artificial Consciousness*. A. Chella and R. Manzotti. Exeter, Imprint Academic.

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

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

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

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

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

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

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

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

Erickson, G. J. and C. R. Smith, Eds. [1988] *Maximum-Entropy and Bayesian Methods in Science and Engineering*. Dordrecht, Kluwer.

Izquierdo, E. and E. Di Paolo [2005] Is an Embodied System Ever Purely Reactive? *Proceedings of the 8th European Conference on Artificial Life*. M. S. Capcarrere, A. A. Freitas, P. J. Bentley, C. G. Johnson and J. Timmis. Berlin, Springer-Verlag, 252-261.

Jaynes, E. T. [2003] Probability Theory: The Logic of Science (Cambridge, CUP).

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

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

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

Shannon, C. E. [1948] A mathematical theory of communication, *Bell System Technical Journal* **27**, 379-423 and 623-656, http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.

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

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

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