NEUROSCIENCE of Consciousness

Neuroscience of Consciousness, 2017, 1–9

doi: 10.1093/nc/nix011
Opinion Paper

Integrated information and dimensionality in continuous attractor dynamics

Satohiro Tajima[1,2,*] and Ryota Kanai[3]

[1]Département des Neurosciences Fondamentales, University of Geneva, CMU, rue Michel-Servet 1, Genève, 1211, Switzerland; [2]JST PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama, 332-0012, Japan and; [3]ARAYA, 2-8-10 Toranomon, Minato-ku, Tokyo, 105-0001, Japan

*Correspondence address: CMU, Université de Genève, rue Michel-Servet 1, Genève 1211, Switzerland. E-mail: satohiro.tajima@gmail.com

Abstract

There has been increasing interest in the integrated information theory (IIT) of consciousness, which hypothesizes that consciousness is integrated information within neuronal dynamics. However, the current formulation of IIT poses both practical and theoretical problems when empirically testing the theory by computing integrated information from neuronal signals. For example, measuring integrated information requires observing all the elements in a considered system at the same time, but this is practically very difficult. Here, we propose that some aspects of these problems are resolved by considering the topological dimensionality of shared attractor dynamics as an indicator of integrated information in continuous attractor dynamics. In this formulation, the effects of unobserved nodes on the attractor dynamics can be reconstructed using a technique called delay embedding, which allows us to identify the dimensionality of an embedded attractor from partial observations. We propose that the topological dimensionality represents a critical property of integrated information, as it is invariant to general coordinate transformations. We illustrate this new framework with simple examples and discuss how it fits with recent findings based on neural recordings from awake and anesthetized animals. This topological approach extends the existing notions of IIT to continuous dynamical systems and offers a much-needed framework for testing the theory with experimental data by substantially relaxing the conditions required for evaluating integrated information in real neural systems.

Key words: theories and models; consciousness; computational modeling; dynamical systems; topology; complexity

Introduction

There is a growing interest in the integrated information theory (IIT) of consciousness. The central hypothesis of IIT is that consciousness is integrated information within collective neuronal dynamics (Tononi 2004, 2008; Balduzzi and Tononi 2008; Oizumi *et al.* 2014; Tononi *et al.* 2016). An attractive aspect of IIT is that it could relate basic properties of subjective experience to the physical mechanisms of biological (and even artificial) dynamical systems via an information theoretic framework (Tononi *et al.* 2016). In particular, IIT is based on partitioning the system, and these partitions reveal irreducible sets of elements in the system; those elements or parts of the system corresponding to the conscious experience. Among the theories of consciousness, IIT is relatively new and still awaits empirical verification. To examine IIT with empirical neural recordings, however, its current implementation needs to address several issues from both practical and theoretical viewpoints. Although empirical studies have reported neural phenomena for which IIT could provide consistent explanations (Massimini *et al.* 2005, 2007; Lee *et al.* 2009; Casali *et al.* 2013; Sasai *et al.* 2016), it is still challenging to test the necessity of IIT directly with empirical datasets under its current formulation. For example, measuring integrated information in a rigorous sense requires observing all the elements at the same time, which imposes a serious bottleneck to

Received: 6 February 2017; **Revised:** 13 April 2017. **Accepted:** 21 April 2017

testing the theory with neural recordings in living organisms. In addition, deriving exact values of the integrated information is often computationally intractable for systems with a large number of elements.

Here, we discuss an alternative implementation of IIT that could resolve some aspects of those problems. A key idea in our formulation is to index the integrated information in terms of the topological dimensionality of shared attractor dynamics. In this formulation, the effects of unobserved nodes on the attractor dynamics can be reconstructed using a technique called *delay embedding* (Takens 1981; Sauer *et al.* 1991). This technique allows us to reconstruct the properties of global multivariate states from time series observed in the subset of variables. As we will discuss later in this article, such reconstructability from partial observation also relates to a conceptual issue: that the effects of spatial partitioning may depend on how time sequences are chunked to define the momentary "states". Remarkably, considering topological properties allows us to make use of, rather than suffer from, such a puzzling property of reconstructability in continuous dynamical systems. We illustrate how this formulation works with simple examples and discuss its relevance to the original formulation of IIT and our recent empirical findings from awake and anesthetized animals (Tajima *et al.* 2015).

The aim of the present Opinion Paper is to illustrate the basic idea behind our formulation. For this purpose, we focus on intuitive rather than rigorous mathematical descriptions. Additionally, the interpretation of the "integrated information" depends on the version of IIT. For example, the latest framework of IIT (so-called "IIT 3.0"; Oizumi *et al.* 2014) requires more steps to assess the integrated information than in previous version ("IIT 2.0"; e.g., Balduzzi and Tononi 2008). In this article, to keep the arguments simple and accessible to the general readership, we basically focus on the definition of integrated information introduced in IIT 2.0 (which is more similar to the "small-phi" (ϕ) rather than "big-phi" (Φ) in IIT 3.0).

Topological Dimensionality as an Indicator of Integrated Information in Continuous Dynamical Systems

To illustrate our formulation, let us consider simple dynamical systems consisting of only two nodes, with values of x_1 and x_2 (**Fig. 1**). Suppose that each of x_1 and x_2 has self-feedback, which is generally nonlinear. For the sake of simplicity, here we assume that each node's value is defined in one-dimensional continuous space (e.g., $x_1,\ x_2 \in \mathbb{R}$), but the subsequent arguments are valid for general cases in which node values are defined in higher-dimensional spaces.

First, let us begin by considering a mutually interacting system (**Fig. 1a–i**). Since similar arguments apply to the cases with ordinary differential equations, assume the system dynamics to be described with deterministic difference equations as:

$$x_1^t = f(x_1^{t-1}, x_2^{t-1}), \tag{1}$$

$$x_2^t = g(x_1^{t-1}, x_2^{t-1}), \tag{2}$$

where f and g are arbitrary continuous functions and t denotes an arbitrary time point. In a general nonlinear system, the state (x_1^t, x_2^t) could be distributed across at most a 2-dimensional manifold A (the light gray square in the figure) in the phase space of $(x_1,\ x_2)$. For convenience, we call A an "attractor" when the state stays within A for a sufficiently long time. After a sufficient duration away from the initial state, attractor A provides some joint probability density distribution $p(x_1,\ x_2)$.

Suppose that we could identify both nodes' values (x_1^t, x_2^t) at time t by observing them simultaneously (i.e., we could make the joint probability density distribution $p(x_1^t, x_2^t)$ be a delta function through the observation). If we consider the past state of those nodes, (x_1^{t-1}, x_2^{t-1}), based on this observation, the uncertainty in the inference is described by the conditional distribution $p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t)$. In a general nonlinear deterministic system, $p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t)$ is supported by a finite number of points in attractor A, except for some special cases (like either function f or g being flat). To rephrase, identifying the system's current state (x_1^t, x_2^t) constrains its past state (x_1^{t-1}, x_2^{t-1}) on a set of zero-dimensional manifolds (i.e., points). **Figure 1a–ii** depicts a simple case in which the previous state is perfectly constrained to a single point.

What if we did not use the joint observation, $p(x_1^t, x_2^t)$ but rather inferred the past values of individual nodes separately based on marginal observations, $p(x_1^t)$ and $p(x_2^t)$? Because we assumed no uncertainty in observing each node's current value, $p(x_1^t, x_2^t)$, $p(x_1^t)$ and $p(x_2^t)$ are all delta functions, and we thus have $p(x_1^t, x_2^t) = p(x_1^t)p(x_2^t)$. What about the inferred past state, $p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t)$? In fact, such an equality does not hold between the past state distributions, $p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t)$ and $p(x_1^{t-1}|x_1^t)p(x_2^{t-1}|x_2^t)$. Indeed, the previous state generally cannot be identified by the marginal observation when the nodes interact with each other. Namely, $p(x_i^{t-1}|x_i^t)$ is not described by a delta function of x_i^{t-1} $(i = 1, 2)$, even if $p(x_i^t)$ is a delta function, and thus generally

$$p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t) \neq p(x_1^{t-1}|x_1^t)p(x_2^{t-1}|x_2^t). \tag{3}$$

This fact could be understood intuitively as follows: for example, the set of states that fall in the support of the marginal distribution $p(x_1^t)$ is represented by the vertical red line in **Fig. 1a–iv**, reflecting the uncertainty about the current value of x_2^t. This set is generally mapped to an oblique line (or a curve) in the previous time point (the red curve in **Fig. 1a–iii**) due to the interaction between x_1 and x_2. Because we do not know where the actual past state was on this curve, we have 1-dimensional uncertainty for the past state of x_1 when we consider the projection of this curve onto the x_1-axis (as indicated by the non-zero length of the red bar on the horizontal axis in **Fig. 1a–ii, iii**). The same argument applies to x_2, and thus we have another 1-dimensional uncertainty, now for x_2, as shown by the blue bar on the vertical axis. Together, we have a 2-dimensional uncertainty for the past state inferred from the separate ("partitioned") observations $p(x_1^{t-1}|x_1^t)p(x_2^{t-1}|x_2^t)$ in total (as depicted by the dark gray rectangle in **Fig. 1a-iii**). Now, recalling that the inference based on a joint observation, $p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t)$, had 0-dimensional uncertainty, we can understand that $p(x_1^{t-1}|x_1^t)p(x_2^{t-1}|x_2^t)$ generally differs from $p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t)$.

This inequality between the joint and partitioned conditional probability distributions is key for characterizing the integrated information value (ϕ) in IIT (Tononi 2004; Balduzzi and Tononi 2008; Oizumi *et al.* 2014, 2016a,b), and note that basically the same argument applies to the relationships between the current and future states. How to quantify the difference between the joint and partitioned distributions is arbitrary. Roughly speaking, IIT 2.0 used the Kullback–Leibler divergence (Balduzzi and Tononi 2008; Oizumi *et al.* 2016a) and IIT 3.0 the earth mover's distance (EMD) (Oizumi *et al.* 2014) to quantify the differences between the joint and partitioned probability

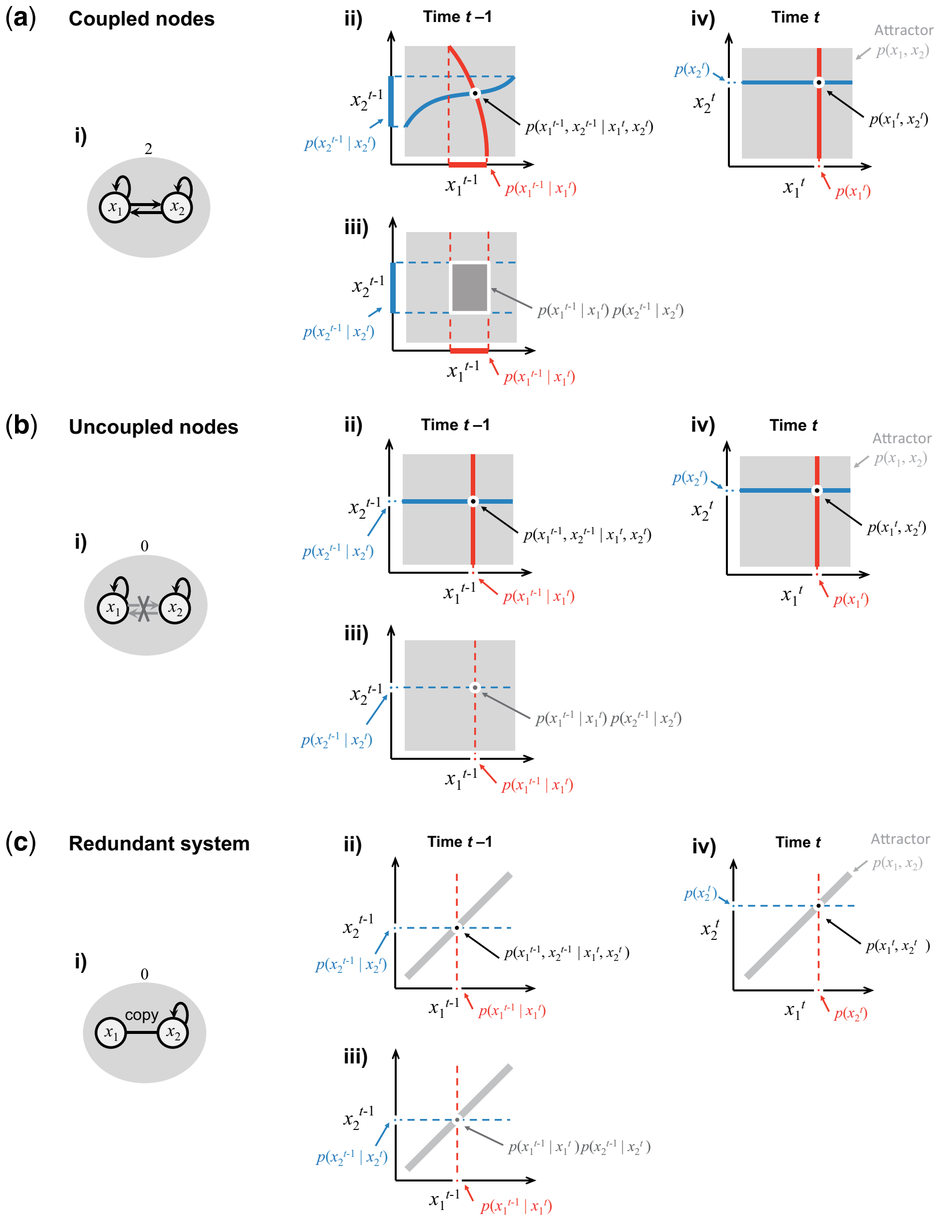

Figure 1. Schematic illustrations for the dimensionality-based index of integrated information. (a) A system with mutually interacting nodes. (b) A system comprising two disconnected nodes. (c) A redundant system, in which a node is a copy of the other node. Insets: (i) The schematic of the systems; (ii) the inferred past states at time $t-1$; (iii) the inferred past state at time $t-1$, based on a partitioned observation; (iv) the current states.

distributions. Other information theoretic indices have been proposed for practical applications (Barrett and Seth 2011; Oizumi *et al.* 2016a; Tegmark 2016).

Here, we propose an alternative way of quantifying the difference in distributions based on the topological dimensionality of uncertainty rather than precise information-theoretic quantities. The idea is simple: for example, in the case we described above, $p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t)$ is supported by a 0-dimensional manifold (i.e., point(s)), whereas $p(x_1^{t-1}|x_1^t)p(x_2^{t-1}|x_2^t)$ was supported by a 2-dimensional manifold (i.e., a rectangle). Then, the difference in terms of topological dimensionality between those two distributions is 2 (as $2-0=2$). Formally, if we denote the dimensionality of the support of a distribution p by $\mathrm{Dim}[p]$, the integrated information in terms of the topological dimensionality (ϕ^{Dim}) can be written as follows:

$$\phi^{\mathrm{Dim}} \equiv \mathrm{Dim}\left[p(x_1^{t-1}|x_1^t)p(x_2^{t-1}|x_2^t)\right] - \mathrm{Dim}\left[p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t)\right]. \quad (4)$$

In this example,

$$\phi^{\mathrm{Dim}} = 2. \quad (5)$$

Throughout this article, we only consider cases in which attractors have integer dimensions, and thus ϕ^{Dim} takes integer values, although we can extend the framework to real values by considering non-integer dimensionality such as fractal dimensions (Mandelbrot 1977; Grassberger and Procaccia 1983). Note that, in contrast to the typical frameworks for IIT, the current dimensionality-based metric works in continuous dynamical systems.

Dimensionality Suggests No Integration in a Mechanistically Disconnected System or a Redundant System

For a sanity check, let us now consider how the dimensionality-based quantification of the integrated information works in a physically/mechanistically separated system, as shown in Fig. 1b–i. This system has self-feedback on each node but no interaction between the nodes:

$$x_1^t = f(x_1^{t-1}), \tag{6}$$

$$x_2^t = g(x_2^{t-1}). \tag{7}$$

Although this system would appear to form an apparently 2-dimensional attractor when we plot the trajectory of (x_1^t, x_2^t), it is actually the product of two smaller dynamical systems. How does this fact affect our measure of integrated information ϕ^{Dim}? As before, while observing the entire system (x_1^t, x_2^t), the possible past state (x_1^{t-1}, x_2^{t-1}) is therefore constrained to a finite number of points (again, except for some special cases). Then, let us say $\mathrm{Dim}\left[p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t)\right] = 0$. What about the partitioned observation, $p(x_1^{t-1}|x_1^t)p(x_2^{t-1}|x_2^t)$? Because there is no interaction between the nodes, both x_1 and x_2 are autonomous in their dynamics. This means that observing x_1 alone provides sufficient information to constrain the past state to be on 0-dimensional manifold(s), and so does observing x_2 alone. Namely, the uncertainty of the partitioned observation $p(x_1^{t-1}|x_1^t)p(x_2^{t-1}|x_2^t)$ is still 0-dimensional, and thus the partition does not increase the dimensionality of the uncertainty. This fact is intuitively represented by the red vertical and blue horizontal lines in **Fig. 1b–ii** (notice the difference from the case in which there are interactions between the nodes; Fig. 1a–ii). Therefore, in this system, $\mathrm{Dim}\left[p(x_1^{t-1}|x_1^t)p(x_2^{t-1}|x_2^t)\right] = 0$, from which we have

$$\phi^{\mathrm{Dim}} = 0. \tag{8}$$

This is the desired result. The above example demonstrates that the dimensionality-based index of integrated information, ϕ^{Dim}, correctly captures the "absence of integration" in a disconnected system, just like in the original formulation of IIT.

Another case in which the integrated information is zero is when the system has apparently two (or more) nodes, but the information is redundant. For an extreme example, if node x_1 is a mere copy of the other node x_2, partitioning the observations of the two nodes does not increase the uncertainty (Fig. 1c), thus $\phi^{\mathrm{Dim}} = 0$ even though the two nodes shows strong apparent coupling in terms of correlation. In general, ϕ^{Dim} is zero when x_1 is solely a function of x_2 (i.e., $x_1 = f(x_2)$) because observing x_1 and x_2 at the same time does not decrease the uncertainty compared to observing x_2 alone. This confirms that not only the coupling among nodes but also the differentiation of the system's states is necessary to have higher ϕ^{Dim} values.

Dimensionality Suggests Integration in an Attractor Reconstructed from Partial Observation

We have seen that the dimensionality-based index of integrated information seems to yield reasonable results for simple examples. However, what are the advantages of considering the dimensionality instead of the precise information quantity? To see this, we now turn to considering a case in which we do not observe the entire system but can access only part of it—say, x_1 alone (Fig. 1). As in the first example, let us assume mutual interaction between the two nodes. In contrast to the previous cases, however, now we assume that we never have access to x_2.

Usually, there is no means of measuring the integrated information between x_1 and x_2 when we cannot observe x_2's state. However, because x_1 and x_2 are interacting, x_2's information could be implicitly coded by the temporal evolution of x_1. If so, we might be able to reconstruct some aspects of the dynamics of the entire system from the observation of its subset. This is indeed the case in nonlinear, deterministic dynamical systems in general—via a mathematical technique known as "delay embedding" (Takens 1981; Sauer *et al.* 1991). Delay-embedding theorems claim that, in short, the temporal pattern of a single variable has a smooth one-to-one mapping to the state of the entire system that the observed variable belongs to. In general, if we have an autonomous dynamical system comprising N variables $(x_1, \ldots, x_N)$ that interact with each other, the trajectory of $(x_1^t, \ldots, x_N^t)$ forms an attractor in this N-dimensional space. Let us say we can only observe the time series of x_1^t. According to the delay-embedding theorems, we can reconstruct the attractor's topology (a shape defined based on connectivity) by plotting the trajectory of $\left(x_1^t, x_1^{t-\tau}, \ldots, x_1^{t-(d-1)\tau}\right)$ instead of $(x_1^t, \ldots, x_N^t)$ when d is sufficiently large, where τ is the unit delay and d is the embedding dimension. It is known that if d is larger than the original attracter's dimensionality, the attractor can be reconstructed almost anywhere on itself with an ignorable volume of overlaps (Sauer *et al.* 1991). It might be somewhat surprising that the property of global dynamics can be reconstructed (in a topological sense) solely from local observation, although it is proven to be the case in almost any type of nonlinear, deterministic dynamical system.

Dimensionality is one of the topological properties reconstructed through the delay embedding. Thus, it is tempting to expect that the present dimensionality-based index of integrated information could be inferred (at least to some extent) from a partial observation. This is indeed possible, with some tweaking, as shown below. Now, let us see how this idea works in our simple example (Fig. 1).

Because we can only observe x_1, we consider a 2-dimensional delay coordinates $(x_1^t, x_1^{t-\tau})$ instead of the original 2-dimensional state space (x_1^t, x_2^t). The unit delay τ could be chosen arbitrarily (or could be optimized in practical data analyses; see "Discussion" section). Let A_1 denote the reconstructed attractor. These 2-dimensinal coordinates are generally not sufficient for embedding when the original attractor A's dimensionality is 2, as the reconstructed attractor A_1 overlaps with itself. Nonetheless, because the original attractor is 2-dimensional, identifying a 2-dimensional state $(x_1^t, x_1^{t-\tau})$ in these delay coordinates can constrain the original state (x_1^t, x_2^t) within a finite set of points (0-dimensional manifold), as long as the number of self-overlaps is finite (which seems to be the case except in pathological situations). This means that we can also infer the past state in the delay coordinates, $(x_1^{t-1}, x_1^{t-\tau-1})$, with 0-dimensional uncertainty.

As the reader may notice, this situation is quite similar to the case in which we could observe the entire system (Fig. 1a). Then, what happens if we consider the partitioned observation, just as before? Now, we cannot consider the spatial partition (because we observe only a single node!). Instead, let us introduce a new partition: a "temporal partition". That is, we consider the partition between temporally distant observations x_1^t and $x_1^{t-\tau}$. Applying the same arguments to this temporal partition reveals that observing either x_1^t or $x_1^{t-\tau}$ alone leads to 1-dimensional uncertainty about the past states x_1^{t-1} and $x_1^{t-\tau-1}$, respectively, and thus the net dimensionality of the uncertainty is 2 (Fig. 1). Therefore, the dimensionality-based index of the integrated information in the delay coordinates is

$$\phi_1^{\mathrm{Dim}} = 2. \tag{9}$$

The suffix "1" indicates that it is based on the delay coordinate reconstruction with node x_1. Again, the result matches that of the case in which we could observe the entire system (Fig. 1a). This fact is interesting because it means that we could reach the same result based on two distinct data: one from the complete observation of the entire system and the other from the partial observation of its subset. In particular, both results match the dimensionality of the attractor in the mutually interacting system we considered here.

One may suspect that this is only a coincidence, but in fact, the dimensionality of the reconstructed attractor generally gives an upper bound of ϕ^{Dim} in the original space. Indeed, when we have a general d_A-dimensional attractor formed by N mutually interacting nodes $(x_1, \ldots,\ x_N)$, the attractor can be reconstructed within d_A-dimensional delay coordinates of node x_i, $\left(x_i^t, \ldots,\ x_i^{t-(d_A-1)\tau}\right)$, allowing d_A-dimensional self-overlaps. As long as the number of self-overlaps is finite, the same argument applies, resulting in 0-dimensiosnal uncertainty in inferring the past state with the joint observation. A temporal (bi-)partition, $\left\{\left(x_i^t, \ldots,\ x_i^{t-k+1}\right),\ \left(x_i^{t-k}, \ldots,\ x_i^{t-(d_A-1)\tau}\right)\right.$ $(\forall k \in \mathbb{N})$, leads to $d_A - k$ and k-dimensional uncertainties for the individual partitioned observations, and thus the net uncertainty turns out to be d_A-dimensional. Together, $\phi_i^{\mathrm{Dim}} = d_A$. On the other hand, ϕ^{Dim} in the original system is upper-bounded by the attractor dimensions, d_A. When $d_A < N$, ϕ^{Dim} could be smaller than d_A. Interestingly, an appropriate projection of the original N-dimensional space to a d_A-dimensional space (e.g., by clustering the nodes and averaging the node values within each cluster) can recover its upper bound, $\phi^{\mathrm{Dim}} = d_A$—which is analogous to the fact that integrated information can be maximized by appropriate coarse-graining (Hoel *et al.* 2013, 2016; Hoel 2016).

Dimensionality Suggests the "Exclusion" of Upstream Nodes in an Asymmetric Interaction

The previous examples show that ϕ^{Dim} in a mutually interacting system reflects the dimensionality of the attractor (whether the observation is over the whole or partial system), whereas $\phi^{\mathrm{Dim}} = 0$ in a disconnected system. Note that the apparent dimensionalities of the attractors were both 2 in those examples (Fig. 1a and b). In this regard, we can interpret ϕ^{Dim} as an index of "interaction-relevant dimensionality" rather than the apparent dimensionality within the original phase space.

The dimensionality-based characterization becomes even less trivial when we consider a system having a hierarchy in terms of the directionality of interactions. To see this, let us consider an example in which the nodes do not mutually interact but rather have a directed interaction (Fig. 3a–c). Now, the node x_2 affects, but is not affected by, node x_1, which can be formally written as

$$x_1^t = f\left(x_1^{t-1}, x_2^{t-1}\right), \tag{10}$$

$$x_2^t = g\left(x_2^{t-1}\right). \tag{11}$$

We can define x_1 as the "downstream" and x_2 as the "upstream" in the system. Note that the upstream node x_2 forms an autonomous dynamical system by itself, whereas the downstream node x_1 belongs to the dynamical system formed by both x_1 and x_2. As we did earlier, we assume that the system has an apparently 2-dimensional attractor in the phase space of $(x_1,\ x_2)$.

Let us first consider the simultaneous observation of the entire system (Fig. 3a). Applying the same analysis as earlier to this system, we find that identifying the system's current state $\left(x_1^t, x_2^t\right)$ constrains its past state $\left(x_1^{t-1}, x_2^{t-1}\right)$ on a set of zero-dimensional manifolds. On the other hand, in the partitioned observations, identifying the upstream node x_2^t constrains its own past state x_2^{t-1} with 0-dimensinal uncertainty, as it is dynamically autonomous, whereas identifying the downstream node x_1^t leaves a 1-dimensional uncertainty about its past state x_1^{t-1}, reflecting the unknown effect from upstream. Together, the net uncertainty in the partitioned observation is 1-dimensional, and thus the index of integrated information in the system is

$$\phi^{\mathrm{Dim}} = 1. \tag{12}$$

Notably, this value of ϕ^{Dim} under the directed interaction is smaller than under the mutual interaction (Fig. 1a).

What if the observation is partial? There are two possibilities of partial observations: observing only the downstream node x_1 (Fig. 3b) or observing only the upstream node x_2 (Fig. 3c). First, when we observe the downstream alone, we can plot the state trajectory in the delay coordinates $\left(x_1^t, x_1^{t-\tau}\right)$ to reconstruct the topology of the attractor being realized in the entire system $(x_1,\ x_2)$, which has 2 dimensions in this case (Fig. 3b). This situation is the same as that in Fig. 1c, and considering the same temporal partition reveals that the index of integrated information based on this reconstructed attractor is

$$\phi_1^{\mathrm{Dim}} = 2. \tag{13}$$

On the other hand, when we plot a similar trajectory in the 2-dimensional delay coordinates with the upstream $\left(x_2^t, x_2^{t-\tau}\right)$, we can reconstruct only a 1-dimensional manifold (Fig. 3c). Because the reconstructed attractor is 1-dimensional, the temporal partition in this 2-dimensional delay coordinates does not increase the uncertainty, resulting in

$$\phi_2^{\mathrm{Dim}} = 0. \tag{14}$$

To summarize the results presented above, in this system with a directed interaction,

$$\phi_1^{\mathrm{Dim}} > \phi^{\mathrm{Dim}} > \phi_2^{\mathrm{Dim}}. \tag{15}$$

These inequalities illustrate that our dimensionality-based index of integrated information is maximized when it is quantified within the downstream node dynamics, not within the entire system. In particular, the higher integration in a subset of the system rather than the whole demonstrates the axiomatic property of "exclusion" assumed in IIT: namely, the physical substrate of conscious experience has unique borders (e.g., the contents of conscious experience do not include the distinction of one's blood pressure being high or low) (Tononi 2008; Oizumi *et al.* 2014; Tononi *et al.* 2016). In the present framework, it is natural to interpret the system's subset i that maximizes ϕ_i^{Dim} as

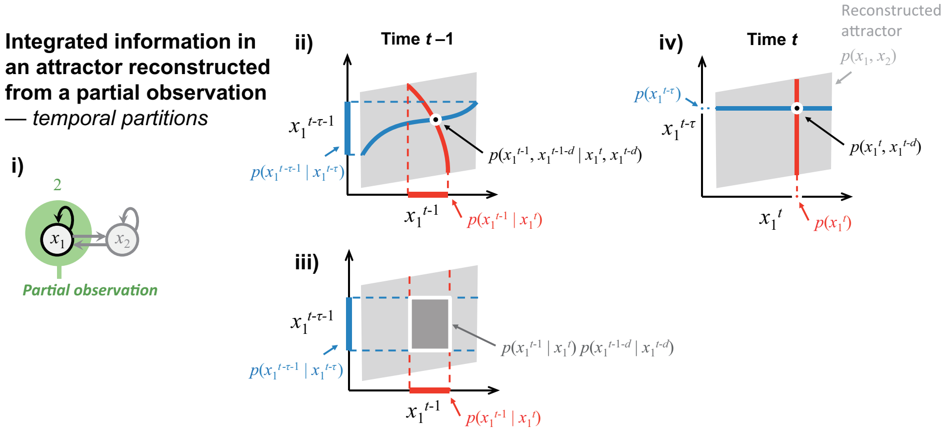
Integrated information in an attractor reconstructed from a partial observation
— temporal partitions
i)
2
x_1
x_2
Partial observation
ii)
Time t –1
$x_1^{t-\tau-1}$
$p(x_1^{t-\tau-1} \mid x_1^{t-\tau})$
$p(x_1^{t-1}, x_1^{t-1-d} \mid x_1^{t}, x_1^{t-d})$
x_1^{t-1}
$p(x_1^{t-1} \mid x_1^{t})$
iv)
Time t
Reconstructed attractor
$p(x_1, x_2)$
$p(x_1^{t-\tau})$
$x_1^{t-\tau}$
$p(x_1^{t}, x_1^{t-d})$
x_1^{t}
$p(x_1^{t})$
iii)
$x_1^{t-\tau-1}$
$p(x_1^{t-\tau-1} \mid x_1^{t-\tau})$
$p(x_1^{t-1} \mid x_1^{t})\, p(x_1^{t-1-d} \mid x_1^{t-d})$
x_1^{t-1}
$p(x_1^{t-1} \mid x_1^{t})$

Figure 2. A partial observation of the system with mutually interacting nodes (the same system as in panel (a)). The inset conventions follow those of Fig. 1.

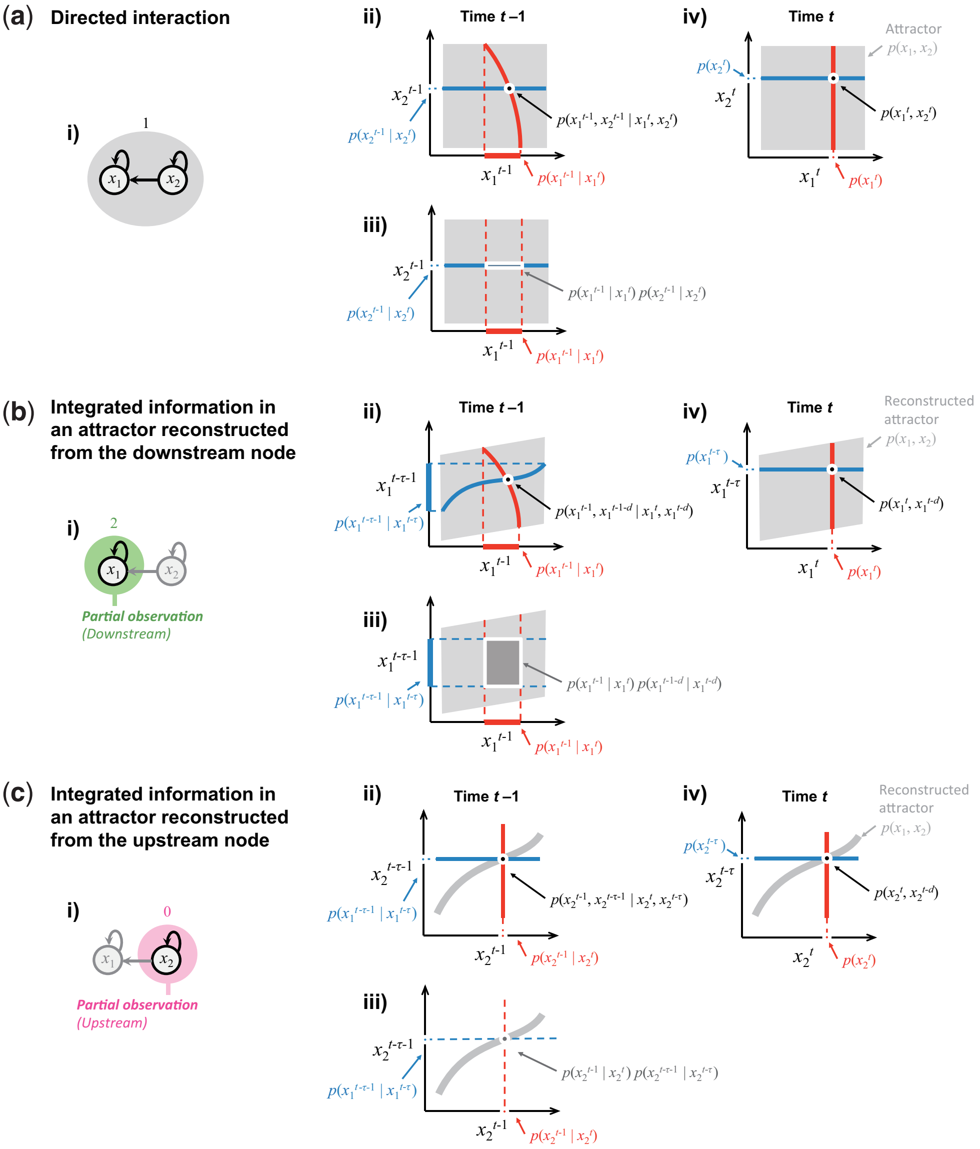
(a) Directed interaction
i)
1
x_1
x_2
ii)
Time t –1
x_2^{t-1}
$p(x_2^{t-1} \mid x_2^{t})$
$p(x_1^{t-1}, x_2^{t-1} \mid x_1^{t}, x_2^{t})$
x_1^{t-1}
$p(x_1^{t-1} \mid x_1^{t})$
iv)
Time t
Attractor
$p(x_1, x_2)$
$p(x_2^{t})$
x_2^{t}
$p(x_1^{t}, x_2^{t})$
x_1^{t}
$p(x_1^{t})$
iii)
x_2^{t-1}
$p(x_2^{t-1} \mid x_2^{t})$
$p(x_1^{t-1} \mid x_1^{t})\, p(x_2^{t-1} \mid x_2^{t})$
x_1^{t-1}
$p(x_1^{t-1} \mid x_1^{t})$
(b) Integrated information in an attractor reconstructed from the downstream node
i)
2
x_1
x_2
Partial observation
(Downstream)
ii)
Time t –1
$x_1^{t-\tau-1}$
$p(x_1^{t-\tau-1} \mid x_1^{t-\tau})$
$p(x_1^{t-1}, x_1^{t-1-d} \mid x_1^{t}, x_1^{t-d})$
x_1^{t-1}
$p(x_1^{t-1} \mid x_1^{t})$
iv)
Time t
Reconstructed attractor
$p(x_1, x_2)$
$p(x_1^{t-\tau})$
$x_1^{t-\tau}$
$p(x_1^{t}, x_1^{t-d})$
x_1^{t}
$p(x_1^{t})$
iii)
$x_1^{t-\tau-1}$
$p(x_1^{t-\tau-1} \mid x_1^{t-\tau})$
$p(x_1^{t-1} \mid x_1^{t})\, p(x_1^{t-1-d} \mid x_1^{t-d})$
x_1^{t-1}
$p(x_1^{t-1} \mid x_1^{t})$
(c) Integrated information in an attractor reconstructed from the upstream node
i)
0
x_1
x_2
Partial observation
(Upstream)
ii)
Time t –1
$x_2^{t-\tau-1}$
$p(x_1^{t-\tau-1} \mid x_1^{t-\tau})$
$p(x_2^{t-1}, x_2^{t-\tau-1} \mid x_2^{t}, x_2^{t-\tau})$
x_2^{t-1}
$p(x_2^{t-1} \mid x_2^{t})$
iv)
Time t
Reconstructed attractor
$p(x_1, x_2)$
$p(x_2^{t-\tau})$
$x_2^{t-\tau}$
$p(x_2^{t}, x_2^{t-d})$
x_2^{t}
$p(x_2^{t})$
iii)
$x_2^{t-\tau-1}$
$p(x_1^{t-\tau-1} \mid x_1^{t-\tau})$
$p(x_2^{t-1} \mid x_2^{t})\, p(x_2^{t-\tau-1} \mid x_2^{t-\tau})$
x_2^{t-1}
$p(x_2^{t-1} \mid x_2^{t})$

Figure 3. Heterogeneity of dimensionality-based index of integrated information under a directed interaction. (a) The index derived based on the observation of the entire system. (b) The index derived based on the observation of the downstream node. (c) The index derived based on the observation of the upstream node. The inset conventions follow those of Fig. 1.

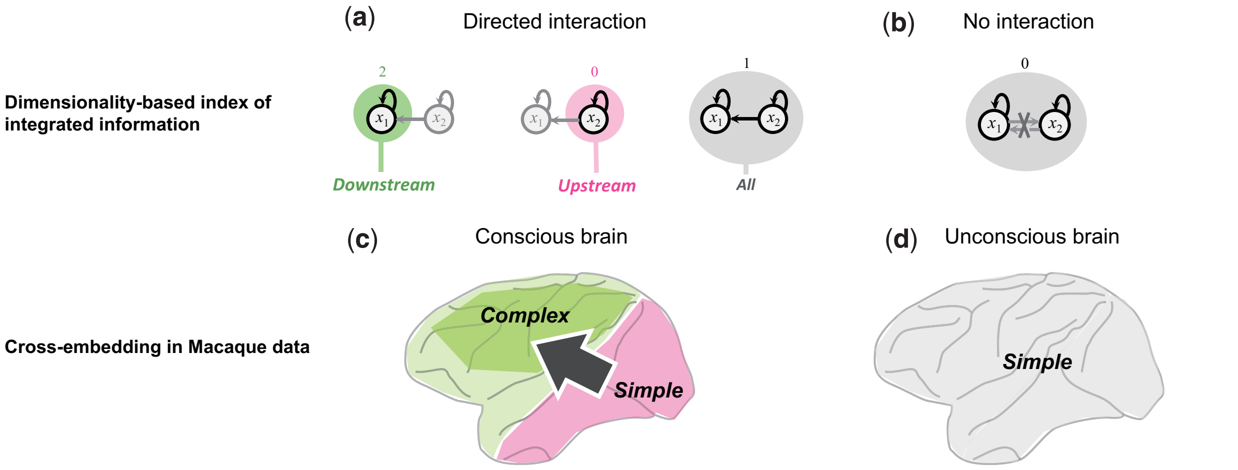

Figure 4. Comparison of the dimensionality-based index of integrated information and the interaction-relevant attractor dimensionality ("complexity") revealed by cross-embedding in conscious and unconscious animals. (a) The system with a directed interaction between nodes (the same as in Fig. 2a–c). (b) The system with no interaction (the same as in Fig. 1b). (c, d) Summary figures modified from Ref. (Tajima *et al.* 2015). (c) The distribution of the attractor complexity revealed by a cross-embedding analysis in awake (conscious) macaque monkeys. (d) The distribution of the attractor complexity revealed by a cross-embedding analysis in anesthetized (unconscious) macaque monkeys.

an analogue of "complex" in IIT, which determines the borders of subjective experiences. Note that the complex in IIT 3.0 (Oizumi *et al.* 2014) is defined with "big-phi" (ϕ), with which systems connected by a purely unidirectional relationship result in "zero" integrated information. Although it is beyond this article's scope to discuss the detailed relations in terms of IIT 3.0, considering big-phi or its proxy would be important for determining the spatial boundaries of physical substrate of consciousness in empirical data (see also "Discussion" section).

Discussion

In this article, we have introduced a dimensionality-based index of integrated information. An advantage of this approach is that it is applicable to continuous nonlinear deterministic systems in general. This complements the existing practical measures that in many cases make use of empirical probability distributions while assuming binary states or linear dynamics with a stochastic component, often for the sake of computational tractability. Our dimensionality-based measure is much simpler than even the original IIT's formulation (Tononi 2004) (Tononi 2004), thus it provides an appealing proxy for integrated information in empirical studies.

Relevance to the Cross-Embedding Complexity

We have seen that our dimensionality-based index of integrated information reflects the dimensions of attractors in dynamical systems in a way sensitive to how the nodes in the systems interact with each other. For example, the value of ϕ^{Dim} can generally differentiate between a mutual interaction (Fig. 1a), a directed interaction (Fig. 3a), and no interaction (Fig. 1b). An alternative way to quantify the interaction-relevant dimensionality (complexity) of the attractor dynamics is "cross-embedding" (Tajima *et al.* 2015). Cross-embedding measures the embedding dimensions necessary for inferring a node's value from the temporal pattern of another node's value by reconstructing attractors in the delay coordinates of individual nodes. Interestingly, analogous to the present dimensionality-based measure of integrated information in interacting and disconnected systems (Fig. 4a and b), the cross-embedding indixes higher dimensionality for the downstream nodes than the upstream nodes. Such a hierarchy of dimensionality is observed in artificial systems having asymmetric interactions as well as in the actual brain dynamics in conscious animals, but much weaker in unconscious animals (Fig. 4c and d) (Tajima *et al.* 2015). Indeed, both the cross-embedding and the dimensionality-based integrated information share the basic idea that high-dimensional attractor dynamics are relevant for consciousness. Moreover, similar to the cross-embedding (Tajima *et al.* 2015), the present index of integrated information demonstrates that information about other nodes can be reconstructed from local dynamics through the delay-embedding technique. This non-localized nature of integrated information can be interpreted as a form of information "broadcasting" among nodes, which the Global Neuronal Workspace Theory associates with consciousness (Dehaene and Changeux 2011). Future studies will investigate more detailed relationships between the cross-embedding and the dimensionality-based integrated information with theoretical analysis and neural recordings.

Spatial Partitions and Coordinate Transformations

To assess the dimensionality of dynamics with partial observations, we introduced the idea of "temporal partitioning" of the attractor reconstructed within delay coordinates, based on the delay-embedding theorems. This is a key contribution of this study that could bridge IIT's framework to empirical data, in which we often have access to only partial observations of the studied system. At the same time, the delay embedding and temporal partitioning may invite a new question about the meaning of the spatial partitions considered in the original formulations of IIT: because we can extract information about unobserved variables through delay embedding by regarding the temporal pattern in a subset of the system as a "state", the conclusion derived from the spatial partition could be affected severely by the definition of state within each node. As it has previously been proven that the spatiotemporal coarse-graining affects both effective information and integrated information (Hoel *et al.* 2013, 2016;), this indicates an avenue of future research may be in how the information leveraging by the delay embedding changes the net integrated information. If so, the present dimensionality-based indexing of integrated information allows us to make use of, rather than suffer from, the effects of embedding in continuous dynamical systems. Moreover, the dimensionality-based assessment could be robust to changes in the definition of states because the topological dimensionality is in many cases invariant to coordinate transformations, even a partial observation of a system. Note that this invariance is gained in exchange of a more detailed characterization of the information-theoretic quantity; the topological dimensionality is a much coarser measure than the usual measures of integrated information due to the topological invariance.

Finding the Complex with Empirical Data

When relating IIT's predictions to recorded neural activities, identifying the "complex" (the set of neurons corresponding to the conscious experience) in the brain will be a crucial step (Koch *et al.* 2016; Tononi *et al.* 2016). A possible description of the "complex" in our framework would be: "the set of elements sharing a maximally irreducible and high-dimensional attractor". Practically, to specify such a set of elements in empirical data would require solving at least two technical issues: inferring whether the observed variables actually share an attractor, and identifying an appropriate spatiotemporal scale to define such dynamics.

Although we have focused on illustrating our present framework with simple examples, in practice it will be a critical problem whether the observed variables belong to the same complex or not, particularly when we analyze real data. One of the possible approaches to identifying complexes is to look at the similarity of measured dimensionality: if two nodes belong to the same complex, they should share the same (or nearly the same, in practice) phi-dim. Another way is to directly assess the causal couplings between the considered nodes by testing the existence of embedding (a one-to-one mapping) between the attractors reconstructed from those nodes ("cross-embedding" used in Tajima *et al.* 2015). This is because, if the two nodes belong to the same complex, they should share the same attractor dynamics, thus, there should generally be a one-to-one relationship between the reconstructed attractors.

In practical data analyses, we also often need to search for an appropriate length of unit delay, particularly when dealing with limited amount of data. In such cases, the efficiency of attractor reconstruction will depend on the dominant timescale (e.g., how slowly the autocorrelation decays) in the observed signals: generally, the longer unit delay works better for the slower dynamics data, because using a longer unit delay effectively means putting more emphasis on slower components than faster ones in the analysis with delay-coordinate reconstructions. In this sense, the choice of a particular unit delay corresponds roughly to analyzing the interactions or integrated information at a particular temporal scale (Hoel *et al.* 2013, 2016). Practically, we could choose the unit delay a priori based on the autocorrelation or mutual information (Fraser and Swinney 1986), or could choose it by directly looking at the resulting ϕ^{Dim} values (e.g., by maximizing ϕ^{Dim}). Although the present article focuses mostly on the idealized situation with a large data size limit, how to select an appropriate spatiotemporal scale will be an important issue when we apply the theory to analyzing real systems.

Limitations

Although we believe that the present topology-based approach will provide both practical and theoretical insights to IIT, there is still room for elaboration. A major limitation of the current framework is that it assumes we can estimate the exact dimensionality of attractors, which can be challenging in real data. To implement the computation described here requires an efficient algorithm for estimating the underlying the attractor dimensionality, however, we expect that it should be possible by extending a dimensionality estimation algorithm similar to the one used in the cross-embedding method (Tajima *et al.* 2015).

Another caveat of the current embedding-based argument is that mathematically rigorous claims can be applied only to continuous, deterministic systems. However, when we consider a grid (or similar form of) representation of states that span some volume in a continuous state space, we can naturally define non-zero dimensions. For example, by considering a version of box-counting dimensions with box sizes larger than the grid intervals. In respect to the determinism, empirical studies with artificial and real data have shown that delay embedding works even in dynamical systems with some stochasticity (Sugihara and May 1990; Sugihara *et al.* 2012; Tajima *et al.* 2015; Ye and Sugihara 2016), although future theoretical studies are required for more thorough verifications of the method. Note also that in realistic situations including stochastic dynamics, some information could be lost in the communications among nodes due to noise or other constraints on signal transmissions (e.g., narrow-band temporal frequency responses) that make the downstream information degenerate. In such cases, the attractor dimensions are not always maximized at the system's downstream as in the examples we discussed—which agrees with our intuition that the maximally integrated information should be observed in the central nervous system, rather than its peripheral downstream (e.g., muscles).

Lastly, although the present study focused on the attractor dimensions and relating them to the integrated information as a measure of the level of consciousness, it remains to be investigated how we can characterize the quality (or contents) of consciousness within this topological framework. A potentially useful approach to characterizing the quality of consciousness is to look at more detailed structures of the attractors, such as the number of holes in each dimension or the higher-order relationships among multiple reconstructed attractors.

Conclusion

Currently, the value of IIT is still a subject of debate, attracting both enthusiasm and criticism (Cerullo 2015). An important next step will be to test the fundamental concepts of IIT empirically. For practical and theoretical reasons, however, it has been difficult to perform a rigorous computation of integrated information from real neural data. Our present study offers one practical measure of the integrated information from real neural data in which the observations are partial and the variables are continuous. Specifically, we have shown that in continuous attractor dynamics, the topological dimensionality of a reconstructed attractor can be used to measure the degree of integrated information. We believe that this captures a critical aspect of integrated information as it is invariant to general coordinate transformations. This topological dimensionality-based characterization is not only consistent with the existing framework of IIT, but it also significantly relaxes the conditions required for evaluating the integrated information. As such, the topological dimensionality enables us to assess the integrated information even from partial observations and provides a much-needed framework for testing the theory with experimental data.

Conflict of interest statement. None declared.

References

Balduzzi D, Tononi G. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLOS Comput Biol* 2008;**4**:e1000091.

Barrett AB, Seth AK. Practical measures of integrated information for time-series data. *PLOS Comput Biol* 2011;**7**:e1001052.

Casali AG, Gosseries O, Rosanova M, *et al*. A theoretically based index of consciousness independent of sensory processing and behavior. *Sci Transl Med* 2013;**5**:198ra105–198ra105.

Cerullo MA. The problem with phi: a critique of integrated information theory. *PLoS Comput Biol* 2015;**11**:1–12.

Dehaene S, Changeux J-P. Experimental and theoretical approaches to conscious processing. *Neuron* 2011;**70**:200–27.

Fraser AM, Swinney HL. Independent coordinates for strange attractors from mutual information. *Phys Rev A* 1986;**33**:1134–40.

Grassberger P, Procaccia I. Characterization of strange attractors. *Phys Rev Lett* 1983;**50**:346–9.

Hoel EP. When the map is better than the territory. 2016; *arXiv*:1612.09592.

Hoel EP, Albantakis L, Marshall W, *et al*. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neurosci Conscious* 2016;**1**:1–13.

Hoel EP, Albantakis L, Tononi G. Quantifying causal emergence shows that macro can beat micro. *Proc Natl Acad Sci* 2013;**110**:19790–5.

Koch C, Massimini M, Boly M, *et al*. Neural correlates of consciousness: progress and problems. *Nat Rev Neurosci* 2016;**17**:307–21.

Lee U, Mashour GA, Kim S, *et al*. Propofol induction reduces the capacity for neural information: implications for the mechanism of consciousness and general anesthesia. *Conscious Cogn* 2009;**18**:56–64.

Mandelbrot BB, (1977) *Fractals—Form, Chance and Dimension*. San Francisco: Freeman.

Massimini M, Ferrarelli F, Esser SK, *et al*. Triggering sleep slow waves by transcranial magnetic stimulation. *Proc Natl Acad Sci USA* 2007;**104**:8496–501.

Massimini M, Ferrarelli F, Huber R, *et al*. Breakdown of cortical effective connectivity during sleep. *Science* 2005; **309**:2228–32.

Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLOS Comput Biol* 2014;**10**:e1003588–1. 25.

Oizumi M, Amari SI, Yanagawa T, *et al*. Measuring Integrated Information from the Decoding Perspective. *PLoS Comput Biol* 2016a;**12**:e1004654:1–18.

Oizumi M, Tsuchiya N, Amari S. A unified framework for information integration based on information geometry. *Proc Natl Acad Sci USA* 2016b;**113**:14817–22.

Sasai S, Boly M, Mensen A *et al*. Functional split brain in a driving/listening paradigm. *Proc Natl Acad Sci* 2016;**113**:14444–9.

Sauer T, Yorke JA, Casdagli M. Embedology. *J Stat Phys* 1991;**65**:579–616.

Sugihara G, May R. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature* 1990;**344**:734–41.

Sugihara G, May R, Ye H, *et al*. Detecting causality in complex ecosystems. *Science* 2012;**338**:496–500.

Tajima S, Yanagawa T, Fujii N, *et al*. Untangling brain-wide dynamics in consciousness by cross-embedding. *PLOS Comput Biol* 2015;**11**:e1004537.

Takens F. Detecting strange attractors in fluid turbulence. In: Rand DA, Young L-S (eds), *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*. Berlin: Springer, 1981, 366–381.

Tegmark M. Improved measures of integrated information. *PLoS Comput Biol* 2016;**12**:1–34.

Tononi G. An information integration theory of consciousness. *BMC Neurosci* 2004;**5**:1–22. 42:

Tononi G. Consciousness as integrated information: a provisional manifesto. *Biol Bull* 2008;**215**:216–42.

Tononi G, Boly M, Massimini M, *et al*. Integrated information theory: from consciousness to its physical substrates. *Nat Rev Neurosci* 2016;**17**:450–61.

Ye H, Sugihara G. Information leverage in interconnected ecosystems: overcoming the curse of dimensionality. *Science* 2016;**353**:922–5.