Neuroscience of Consciousness, 2016, 1–13

doi: 10.1093/nc/niw012
Research article

Can the macro beat the micro? Integrated information across spatiotemporal scales

Erik P. Hoel, Larissa Albantakis, William Marshall and Giulio Tononi*

Department of Psychiatry, University of Wisconsin, Madison, 6001 Research Park Blvd, WI 53703, USA

*Correspondence address. Department of Psychiatry, University of Wisconsin, University of Wisconsin - Madison, 6001 Research Park Blvd, Madison, WI, 53719. Tel: 1-608-263-6063; Fax: 1-608-26502953; E-mail: gtononi@wisc.edu

Abstract

Causal interactions within complex systems such as the brain can be analyzed at multiple spatiotemporal levels. It is widely assumed that the micro level is causally complete, thus excluding causation at the macro level. However, by measuring effective information—how much a system's mechanisms constrain its past and future states—we recently showed that causal power can be stronger at macro rather than micro levels. In this work, we go beyond effective information and consider additional requirements of a proper measure of causal power from the intrinsic perspective of a system: composition (the cause–effect power of the parts), state-dependency (the cause–effect power of the system in a specific state); integration (the causal irreducibility of the whole to its parts), and exclusion (the causal borders of the system). A measure satisfying these requirements, called Φ^{Max}, was developed in the context of integrated information theory. Here, we evaluate Φ^{Max} systematically at micro and macro levels in space and time using simplified neuronal-like systems. We show that for systems characterized by indeterminism and/or degeneracy, Φ can indeed peak at a macro level. This happens if coarse-graining micro elements produces macro mechanisms with high irreducible causal selectivity. These results are relevant to a theoretical account of consciousness, because for integrated information theory the spatiotemporal maximum of integrated information fixes the spatiotemporal scale of consciousness. More generally, these results show that the notions of macro causal emergence and micro causal exclusion hold when causal power is assessed in full and from the intrinsic perspective of a system.

Key words: theories and models; consciousness; computational modeling; emergence; philosophy; reductionism

Introduction

The causal structure of physical systems can be analyzed at various spatial or temporal levels, from the most fine-grained micro level to any coarse-grained macro level. For example, the brain can be analyzed, in space, at the level of neurons, neuronal groups, macro-columns, and areas; and in time, over tens, hundreds, and thousands of milliseconds (Sporns *et al.*, 2005). Practical considerations, lack of detailed data, and heuristic strategies usually dictate the spatiotemporal scale at which a system's causal structure is actually studied, which is often very coarse-grained. Thus, neuroimaging studies of effective connectivity in the brain examine interactions at the spatial level of voxels, which contain millions of neurons, and at the temporal level of blood-oxygen fluctuations, on the order of seconds. While such coarse-grained investigations are useful, it is widely assumed that the causal structure of a system is only fully captured by the most fine-grained causal model. This "micro" assumption is ubiquitous in science and underlies ambitious programs that aim at collecting and modeling data at the finest scale possible (Markram, 2006).

At a theoretical level, this reductionist "micro" bias is based on the following assumptions (Kim, 1993): (i) once the properties of micro-level physical mechanisms are fixed, macro-level properties are fixed too (supervenience); (ii) causal power resides fully at the microphysical level (micro causal closure); and (iii) if all the causal work is done at the micro level, there is no room for any causal contribution at the macro level (macro causal exclusion). This view of causal power denies the possibility of genuine causal emergence. It also rules out any sort of "mental causation" (Kim, 2000), since consciousness is thought

Received: 19 January 2016; **Revised:** 27 May 2016. **Accepted:** 9 July 2016

to supervene upon its physical substrate (PSC) and its features are undoubtedly "macro" rather than "micro," both spatially and temporally.

In this article, we address the issue of causal emergence and question the "micro" assumption about causal power by resorting to the theoretical framework provided by integrated information theory (IIT). IIT is a theory of consciousness that starts from the essential properties of phenomenal experience and derives the requirements that must be satisfied by its PSC. Specifically, IIT argues that the PSC must be a maximum of intrinsic cause–effect power (Tononi, 2004, 2008, 2012, 2015; Oizumi *et al.*, 2014; Tononi *et al.*, 2016). A direct implication of the theory is that the spatiotemporal grain of the physical elements and intervals constituting the PSC must be the one that maximizes cause–effect power.

The spatiotemporal scale of experience is clearly of a macro kind—an "instant" of experience is on the order of tens to hundreds of milliseconds, rather than microseconds (Bachmann, 2000; Holcombe, 2009). Hence, IIT predicts that the PSC must have maximum intrinsic cause–effect power at the level of macro elements and macro intervals, rather than at the level of micro elements and micro intervals (Tononi, 2004; Marom, 2010; Chalmers, 2013). This prediction implies that the reductionist "micro" assumption with respect to causal power must be wrong, i.e. that "the macro can beat the micro" (Hoel *et al.*, 2013).

The quantitative assessment of cause–effect power is a prerequisite for determining whether and under what conditions the macro can indeed beat the micro. Importantly, IIT provides the conceptual and mathematical tools to fully assess cause–effect power, at least for idealized, simple systems that can be manipulated, observed, and partitioned systematically. In recent work, we provided a first proof of principle that, once causal interactions are quantified, causal power at the macro level can surpass that at the micro level (Hoel *et al.*, 2013). However, this evaluation was done using an average measure of causal interactions within a predefined system taken as a whole. In this article, we aim to establish whether "the macro can beat the micro" if causal power is assessed in full (Oizumi *et al.*, 2014), by considering: (i) the cause–effect power of a system on itself (intrinsic existence, independent of external observers); (ii) the cause–effect power of the system's parts within the system (composition); (iii) the cause–effect power of the specific state the system is in (state-dependency); (iv) the requirement that the cause–effect power of a system must be irreducible to that of its parts (integration); (v) the requirement that cause–effect power must be definite, corresponding to the particular set of elements and spatiotemporal grain that is maximally irreducible (exclusion; exclusion expresses the requirement that elements cannot contribute cause–effect power multiple times to different sets, see "Theory" section). According to IIT, these requirements (postulates) correspond to the causal properties that must be satisfied by the PSC, in turn are derived from the essential phenomenal properties of consciousness (axioms): experience exists intrinsically, is structured, specific, unitary, and definite (Tononi, 2012, 2015; Oizumi *et al.*, 2014).

For simple systems, maxima of irreducible, state-dependent, compositional, intrinsic cause–effect power can be evaluated by measuring the non-negative quantity Φ (integrated information, see "Theory" section). This measure of integrated information has already been applied to classify the causal structure of discrete dynamical systems, such as cellular automata (Albantakis and Tononi, 2015), and to track how the causal structure of simulated organisms, called animats, evolves in a simulated environment (Albantakis *et al.*, 2014). Here, we systematically evaluate Φ for simple, idealized systems considered at coarser or finer spatiotemporal grains, and show that, depending on the structure of a given system, Φ can in principle reach a maximum at a macro level, in space and time, rather than at a micro level ($\Phi^{\mathrm{Macro}} > \Phi^{\mathrm{Micro}}$). Such demonstration constitutes a necessary first step toward a principled account of the spatiotemporal scale of consciousness. Moreover, it provides the theoretical foundation for empirical studies aimed at characterizing the neural elements and intervals constituting the neural substrate of consciousness and at testing a key prediction of IIT. For example, if in the brain a maximum of intrinsic cause–effect power were to obtain at the more macro spatial scale of neuronal groups, rather than at the more micro scale of neurons, IIT would predict that only changes in the average activity of a group of neurons, and not of individual neurons, should make a difference to the content of experience (Tononi *et al.*, 2016).

Theory

IIT

A detailed description of IIT 3.0 can be found in Oizumi *et al.* (2014) and Tononi (2015). In the following, we first outline the IIT 3.0 algorithm to find the physical system with the maximal amount of integrated information (Φ^{Max}). A "physical" system is defined as a set of elements, each of which has at least two states, inputs that can affect its state and outputs that are affected by its state. The elements are "physical" in the sense that each individual element can be perturbed and observed, and sets of elements partitioned. Individual elements that cannot be partitioned, spatially or temporally, are "micro" elements. Given a collection of such micro elements in a particular state, the IIT 3.0 algorithm performs a search across all possible physical systems that can be constituted of these elements, to find the system with maximal Φ. Due to computational constraints, we restrict our examples to those where the micro elements are small collections of binary logic gates.

For a given physical system S, we first perturb the elements of S into all possible states with equal probability (Tononi *et al.*, 1999) and the resulting state transitions are observed. These state transitions define the transition probability matrix (TPM) of the physical system, from which all IIT measures can be derived. Perturbing a physical system into a state means to physically intervene on the system, and to explicitly set its elements into that state. While perturbing the elements of the physical system under consideration, all exogenous elements are treated as background conditions: they are held fixed ("frozen") in their current state (Oizumi *et al.*, 2014). This procedure is akin to the calculus of interventions and the *do(x)* operator introduced by Pearl (2000), to identify causal relationships.

Cause–effect structure of a system

In order to calculate Φ for a given physical system, we must first determine its intrinsic cause–effect structure: how the parts of the system, by being in a specific state, constrain the potential past and future states of the system itself. A part of the system is a set of elements in a state m_t that is a subset of S. To determine the cause–effect structure of a system, we consider every part of the system as a candidate mechanism. It is not necessary for a mechanism to constrain the past and future states of the entire system, it is sufficient that it constrains the state of any subset $Z_{t\pm1}$ of the elements in the system, termed the mechanism's purview. The way a candidate mechanism constrains the potential past and future states of a purview is described by its cause–effect repertoire (Tononi 2015). The cause repertoire is a probability distribution over the states of a past purview (Z_{t-1}) conditioned on the current state of the candidate mechanism, $p_{\text{cause}}(z_{t-1}|m_t)$. Similarly, the effect repertoire is the probability distribution of a future purview (Z_{t+1}) conditioned on the current state of the candidate mechanism, $\text{p}_{\text{effect}}(z_{t+1}|m_t)$, where Z_{t+1} can differ from Z_{t-1}.

To contribute to the intrinsic cause–effect structure of a physical system, the candidate mechanism must be integrated, that is, its cause–effect power must not be reducible to its parts. To assess the irreducibility of a cause or effect repertoire and assess its integrated information (φ), the repertoire is partitioned into two parts, $P = \left\{m_t^{(1)}, Z_{t\pm1}^{(1)}; m_t^{(2)},\ Z_{t\pm1}^{(2)}\right\}$, such that the connections between parts (1) and (2) have been "cut" (injected with noise; Oizumi *et al.*, 2014). The result is a partitioned cause or effect repertoire that is the product of the repertoires for each half of the partition,

$$p^P_{\text{cause}}(z_{t-1}|m) = p_{\text{cause}}\left(z^{(1)}_{t-1}\middle|m_1\right) \times p_{\text{cause}}\left(z^{(2)}_{t-1}\middle|m_2\right),$$

$$p^P_{\text{effect}}(z_{t+1}|m) = p_{\text{effect}}\left(z^{(1)}_{t+1}\middle|m_1\right) \times p_{\text{effect}}\left(z^{(2)}_{t+1}\middle|m_2\right).$$

The difference the partition makes is the distance between the repertoire and the corresponding partitioned repertoire. In IIT, distances D between probability distributions are measured using the earth-mover's distance (EMD; Oizumi *et al.*, 2014), which quantifies the minimum cost of transforming one probability distribution into another (Rubner *et al.*, 2000; Pele and Werman, 2009). For a cause–effect repertoire to be irreducible, all possible partitions must make a difference to the repertoire. The partition that makes the least difference to the repertoire is the minimum information partition (MIP), and the difference it makes defines the irreducibility of the candidate mechanism over its particular purview.

To find the integrated information of a mechanism, we first separately assess the integrated cause and effect information. For the integrated cause information (φ_{cause}), we search over all possible past purviews to identify the maximally irreducible cause repertoire, with its corresponding past purview, also called the core cause of the candidate mechanism. The integrated effect information (φ_{effect}), along with the maximally irreducible effect repertoire and core effect of a candidate mechanism, is defined in an analogous way:

$$\varphi_{\text{cause}}(m_t) \ = \max_{Z_{t-1}} \min_{P} D\left(p_{\text{cause}}(z_{t-1}|m_t),\ p^P_{\text{cause}}(z_{t-1}|m_t)\right),$$

$$\varphi_{\text{effect}}(m_t) = \max_{Z_{t+1}} \min_{P} D\left(p_{\text{effect}}(z_{t+1}|m_t),\ p^P_{\text{effect}}(z_{t+1}|m_t)\right).$$

The overall φ value of the candidate mechanism is the minimum between its φ_{cause} and φ_{effect} values,

$$\varphi(m_t) = \min\left(\varphi_{\text{cause}},\ \varphi_{\text{effect}}\right).$$

In this way, if a candidate mechanism receives input but gives no effective output, or vice versa, its φ is 0, as it has no cause–effect power within the system.

If a candidate mechanism has integrated information $\varphi > 0$ then it is a proper mechanism. The past and future purviews identified by this procedure are its "core cause" and "core effect", and the repertoires over these purviews are its "maximally irreducible cause–effect repertoire". If a mechanism in a state has irreducible cause–effect power (by having $\varphi > 0$), we refer to its maximally irreducible cause–effect repertoire as specifying a "concept". The set of all concepts makes up the cause–effect structure of the physical system $C(S)$. The cause–effect structure of a physical system reveals the compositional nature of the system's intrinsic cause–effect power, identifying not only elementary mechanisms (consisting of a single element), but also irreducible higher-order mechanisms (consisting of multiple elements) that contribute more than the sum of their parts. According to IIT, a concept specifies a phenomenal distinction that contributes to what it is like to be a physical system in its current state.

Cause–effect power of a system

Having obtained the cause–effect structure of a given physical system, the next step in the IIT algorithm is to determine whether the cause–effect structure, as a whole, is irreducible to its parts (integration). If some parts of a physical system make no difference to other parts of the system, then its cause–effect structure is reducible, and the system cannot be a whole from its own intrinsic perspective. To assess the irreducibility of the system, unidirectional bipartitions $P_{\rightarrow} = \left\{S^{(1)}; S^{(2)}\right\}$ of the physical system S are performed, by cutting (injecting with noise) the connections from a subset of elements $S^{(1)}$ to the remaining elements $S^{(2)}$, which creates a partitioned system $S^{P_{\rightarrow}}$ (Oizumi *et al.*, 2014). We then calculate the cause–effect structure of the partitioned physical system $C(S^{P_{\rightarrow}})$, and compare it to $C(S)$ to evaluate the difference made by the partition. A search over all possible directed partitions is performed to identify the one that makes the least difference to the cause–effect structure, its MIP. Integrated information (Φ, "big phi"), measures the irreducibility of a cause–effect structure, by quantifying the difference the MIP makes to the concepts and their φ values of the system,

$$\Phi(S) = \min_{P_{\rightarrow}} D\left(C(S),\ C\left(S^{P_{\rightarrow}}\right)\right).$$

The distance D between two cause–effect structures is assessed using an extended version of the EMD: the cost of transforming one cause–effect structure into another is the amount of φ that needs to be shifted, multiplied by the distance it needs to be moved, where the distance it moves in cause–effect space is the EMD between the concepts' cause–effect repertoires. For full details and examples on the Φ calculation, see Oizumi *et al.* (2014).

The Φ values of all candidate systems are compared to find the maximum (Φ^{Max}). The set of elements with Φ^{Max} is called the "complex". A complex, then, is a physical system that specifies a maximally irreducible cause–effect structure, also called a "conceptual structure". A complex has causal borders and exists as an "intrinsic" entity, from its own intrinsic perspective. This means that its borders are set by its own intrinsic cause–effect structure, as opposed to being set by an external observer. Complexes cannot overlap, as this would imply that the cause–effect power of a shared element would be multiplied for free ("causal exclusion"). According to IIT, there is an identity between the conceptual structure specified by a complex in its current state and its subjective experience—what it is like to be the complex (Oizumi *et al.*, 2014).

Coarse-graining

A discrete, finite system constituted of elements in a state can be considered at various spatiotemporal levels, from the most fine-grained micro level (S_m) to a multitude of coarse-grainings

(S_M). The micro-level S_m is special because, when its features are fully fixed, all its coarse-grainings S_M are also fixed, a property known as “supervenience” (Stalnaker, 1996). In this study, our objective is to identify sets of elements that specify global maxima of integrated information (Φ^{Max}) across elements and spatiotemporal scales. To that end, we extend the algorithmic search for Φ^{Max} across all candidate systems to include all sets of coarse-grained (“macro”) elements across all spatial and temporal levels of the system.

Unless otherwise specified, the micro level is always composed of binary first-order Markov elements {A, B, C …} with possible states {0,1}. For simplicity, without loss of generality, we confine our analysis to coarse-grains in which macro elements are also binary. Macro states will be referred to using {OFF, ON} and macro elements using Greek letters $\{\alpha, \beta, \gamma \ldots\}$. The relationship between the micro level and any of its macro levels can be formalized as a mapping, $\mathbb{M} : S_m \to S_M$. First, one chooses a candidate system S_m of micro elements and its associated bipartitions. Second, disjoint subsets of micro elements from S_m are grouped into macro elements. Third, the associated micro states are mapped into binary macro states. In order to be a valid mapping, S_M must be such that mappings of micro states into macro states are limited to those in which the identity of the individual micro elements within a macro element is irrelevant to determine the macro state (or else the macro level would not be a true coarse-grain as micro-level information would still be available at the macro level; moreover, from the intrinsic perspective of the macro system this information is not available). If, for example, two micro elements are coarse-grained into one macro element, the micro states {01, 10} must be mapped into the same macro state, as distinguishing between them would require information about which micro element is which.

To obtain the TPM of the macro level, S_M must be perturbed into all its possible macro states with equal probability, in the same way as done for the micro level. Perturbing a set of macro elements (setting it to a macro state with the $do(x)$ operator) is done using a macro perturbation, which is the average over perturbations into all n_{micro} micro states that are grouped into the respective macro state s_M:

$$do(S_M = s_M) = \frac{1}{n_{micro}} \sum_{s_m \in s_M} do(S_m = s_m)$$

The TPM of a candidate system is thus assessed independently at each spatiotemporal level: perturbing S_M into all possible macro states with equal probability typically corresponds to a non-uniform distribution of all possible micro perturbations (except if all macro states are composed of the same number of micro states). This reshaping of micro perturbations at the macro level is what makes the causal analysis sensitive to the higher-level causal structure (Hoel *et al.*, 2013). Macro cause–effect structures are then calculated from the macro TPM of a candidate system of macro elements S_M as described above. To be noted, even when coarse-graining, the overall irreducibility Φ of the candidate set is assessed using bipartitions of the micro elements S_m rather than the macro elements S_M. In this case, the goal is to assess, through a “physical” cut of S, the overall cause–effect power of S, above and beyond its “minimal physical” parts.

At the micro and each macro level, the Φ values of all possible candidate systems are evaluated and the system with max(Φ) at each particular level is selected for comparison between levels, yielding the absolute maximum of integrated information (Φ^{Max}) across sets of elements and spatiotemporal scales. For a given macro level, the set of all candidate systems evaluated to optimize Φ includes all possible partitions of micro elements into macro elements of the appropriate scale, as well as all possible mappings between micro states and macro states. If Φ^{Max} is found at a macro level, we can conclude that the system demonstrates “macro causal emergence”. In these cases, $\max(\Phi(S_M)) - \max(\Phi(S_m))$ indicates how much integrated information is gained by analyzing the system at the macro level. Macro causal emergence is spatial, if the winning mapping groups multiple micro elements along with their states into a macro element at a single micro timestep, and temporal if the macro elements consist of only a single micro element but the element’s states are grouped over multiple micro timesteps.

All binary coarse-grains of discrete systems of logic gates were created with a custom-made Python program (PyPhi, see Mayner and Marshall, 2015), available for download at www.integratedinformationtheory.org. (The example systems used here, along with the code required to calculate our results, are available online at https://github.com/wmayner/pyphi.) PyPhi also calculated max(Φ) at each level. Data plots and images were created using MATLAB. Specific examples of spatial and temporal causal emergence are shown below in the “Results” section, along with comparisons of macro concepts to their underlying micro concepts.

Repertoire size, selectivity, and selectivity shift

As outlined above, φ measures the difference that a partition $P = \left\{m_t^{(1)}, Z_{t\pm1}^{(1)}; m_t^{(2)},\ Z_{t\pm1}^{(2)}\right\}$ of mechanism m makes to the mechanism’s cause and effect repertoires over its purviews $Z_{t\pm1}$. To better understand the causal ramifications of coarse-graining, it is helpful to decompose this difference into three components: (i) the repertoire size of cause or effect repertoires, (ii) the change in how selective the mechanism is about its possible causes and effects post-partition, and (iii) the shift in which states are selected as possible causes and effects post-partition (Fig. 1A). The decomposition can be applied to integrated information on both the cause (φ_{cause}) and the effect ($\varphi_{\mathrm{effect}}$) side, but as with integrated information, it is only the side with minimum φ that counts for the mechanism.

The “repertoire size” is the number of elements in the purview. It reflects the degrees of freedom of potential causes or effects; all else being equal higher repertoire size corresponds to higher φ. Note that the definition of repertoire size given here is a special case for systems of binary elements, and that in general, one needs to consider the maximum EMD distance between repertoires, which depends both on the number of elements in the purview and the distance between states of those elements.

The “irreducible selectivity” of a mechanism describes how much the mechanism constrains the past and the future above and beyond its parts. Selectivity of a repertoire can be measured as the difference D (EMD) between the cause or effect repertoire $p(z_{t\pm1}|m_t)$ and the maximum entropy distribution $p(H)$ over all possible states of $Z_{t\pm1}$, normalized by “repertoire size”:

$$selectivity(m_t, Z_{t\pm1}\ ) = \frac{D(p(z_{t\pm1}|m_t),\ \ p(H))}{repertoire\ \ size}.$$

To capture the “irreducible selectivity” of a mechanism above and beyond its parts, we moreover subtract the distance between the partitioned repertoire $p^{MIP}(z_{t\pm1}|m)$ and $p(H)$. In this way:

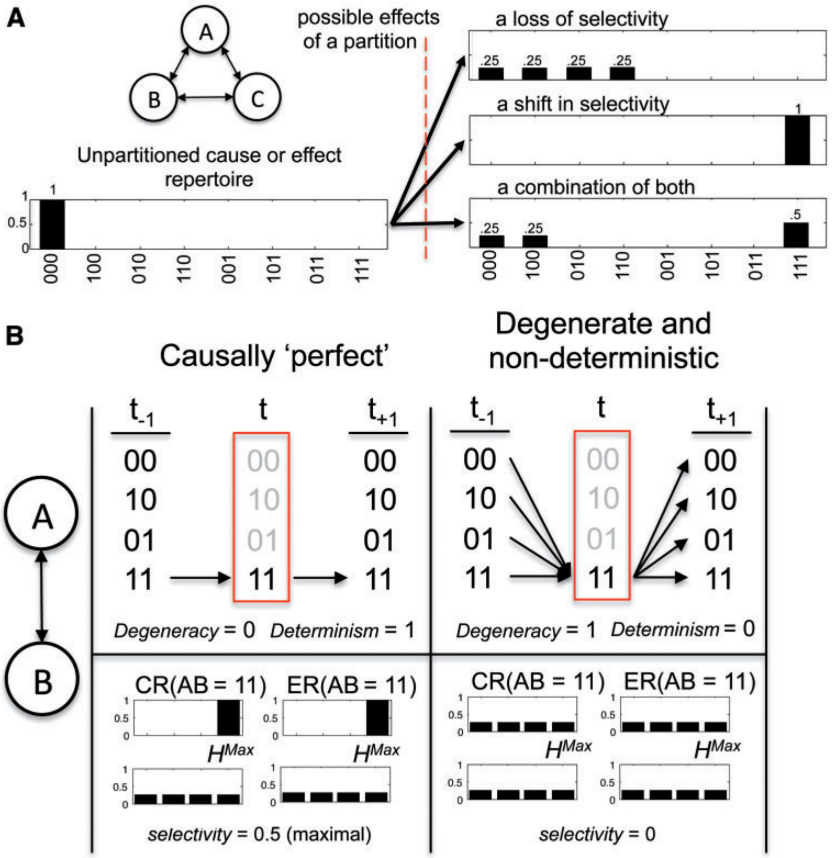


Figure 1. The causal components of integrated information (ϕ). (A) Consider a hypothetical isolated system constituted of three interconnected binary elements. The unpartitioned cause–effect repertoires of the system can change in several ways following a partition. There can be a loss of selectivity, moving the partition closer to maximum entropy (top), a shift in which states are selected in the partitioned repertoire (middle), or a mix of both (bottom). (B) Consider a simpler system of just two connected binary elements (left). If the mechanism AB in state [11] at t could only originate from [11] at t_{-1}, and can only go to [11] at t_{+1}, then degeneracy is 0 and determinism is 1. (B, top left) If AB in state [11] at t could have originated from any state at t_{-1}, and could go to any state at t_{+1}, all with equal probability (B, top right), the mechanism in state [11] has a degeneracy of 0 and a determinism of 1. Compare degeneracy and determinism to selectivity: the minimum distance of either the cause or effect repertoires from the maximum entropy distribution (H). In both cases, selectivity accurately reflects determinism and degeneracy (B, bottom).

$$\mathit{irreducible\ \ selectivity}(m, Z_{t\pm1}) = \frac{(D(p(z_{t\pm1}|m), p(H)) - D(p^{\mathrm{MIP}}(z_{t\pm1}|m),\ \ p(H)))}{\mathit{repertoire\ \ size}}$$

Irreducible selectivity values can range between 0.5 and −0.5, inclusively (negative values are rare but possible if the partitioned cause–effect repertoire is more different from $p(H)$ than the intact one). Selectivity can be related to the notion of determinism and degeneracy as described in (Hoel *et al.*, 2013). There, we demonstrated that the effective information (EI) in a causal model depends on how deterministic and degenerate its mechanisms are on average. "Determinism" (causal divergence) indicates how reliably the current state of a mechanism leads to future states: determinism is 1 when the current state leads to a single future state with probability $p = 1$, and is 0 when all future states have equal probabilities ($p = 1/n$, where n is the number of states).

"Degeneracy" (causal convergence) indicates how many states converge to the same state: degeneracy is 1 when the current state could have come from any previous state with probability $p = 1/n$, and is 0 when the current state could only have come from a single previous state with probability $p = 1$. If determinism = 1 and degeneracy = 0, then that mechanism in a state is causally "perfect" in that it demonstrates maximum selectivity over the states of its purviews (Fig. 1B). If determinism = 0 and degeneracy = 1, then there is a total absence of selectivity over the mechanism's purviews (total noise, Fig. 1B). Previously we showed that it is through increasing the determinism and/or decreasing the degeneracy of causal relationships that coarse-graining can result in higher cause–effect power (Hoel *et al.*, 2013). Since the selectivity of a mechanism M in state m is always positive, it provides the upper limit for the mechanism's irreducible selectivity. If the minimum is $\varphi = \varphi_{\mathrm{Effect}}$, then irreducible selectivity is bounded by the determinism of the mechanism, as the effect repertoire captures divergence into multiple future states. If the minimum is $\varphi = \varphi_{\mathrm{Cause}}$, then irreducible selectivity is bounded by the mechanism's degeneracy, as the cause repertoire captures the convergence from multiple past states (Fig. 1B).

The third component of φ, the "selectivity-shift", captures how the mechanism makes a difference to the system by selecting some specific past or future states over other specific ones (see Supplementary Data S1 for a detailed explanation of "selectivity-shift"). For example, consider a partition that results in 0 irreducible selectivity, but changes which states are specified by the mechanism (Fig. 1A). In this (hypothetical) case, the resulting non-zero φ value is due entirely to a selectivity-shift: the rearrangement of probability mass post-partition without a change in the distance from *p(H)*. Selectivity-shift can be captured by the increase in distance (how much of a "detour" it is), to pass through the partitioned repertoire on the way from the unpartitioned repertoire to $p(H)$, as opposed to going there directly:

$$\begin{aligned}&\textit{selectivity-shift}\\&= \frac{(D(p(z_{t\pm1}|m), p^{\mathrm{MIP}}(z_{t\pm1}|m)) - |D(p^{\mathrm{MIP}}(z_{t\pm1}|m), p(H)) - D(p(z_{t\pm1}|m), p(H))|)}{\textit{repertoire size}}\end{aligned}$$

In summary, as proven in Supplementary Data S2, φ can be fully decomposed into these three quantities: repertoire size, irreducible selectivity, and selectivity-shift:

$$\varphi = (|\textit{irreducible selectivity}| + \textit{selectivity-shift}) * \textit{repertoire size}$$

Results

Finding the maximal value of Φ at a particular spatiotemporal grain requires an algorithmic search across all possible subsets of a given system. Here, this search is expanded to include all possible binary coarse-grains. The figures for each example of macro causal emergence show the result of this expanded search: the macro level (S_M) with Φ^{Max}, along with the micro level (S_m) with highest Φ for comparison.

An illustrative example of spatial coarse-graining and macro causal emergence

To begin with a simple example, consider a four-element system $S_m = \{ABCD\}$ in state [0000], where each micro mechanism operates as an AND gate (with two inputs) under noisy conditions (Fig. 2A). In Hoel *et al.* (2013), this system showed spatial causal emergence in terms of EI. The results of applying the additional causal criteria required by IIT can be seen in Fig. 2B. Each micro element is associated with a single micro concept, for which $\varphi = 0.17$. To visualize the cause–effect structure of the system, each concept is plotted as a star in cause–effect space (Fig. 2C). In cause–effect space, each dimension is a possible past or future state of the system. The four concepts of S_m each occupy a position based on the probability distributions of its maximally irreducible cause–effect repertoires. The size of each star represents how irreducible the concept is (its φ value). Observing the constellation of concepts for the micro-level system, the concepts are small and clustered together in cause–effect space, with the concepts of A/B and C/D overlapping. $\Phi(S_m)$ is only 0.11. The MIP noises the connections from {AC} to {BD}.

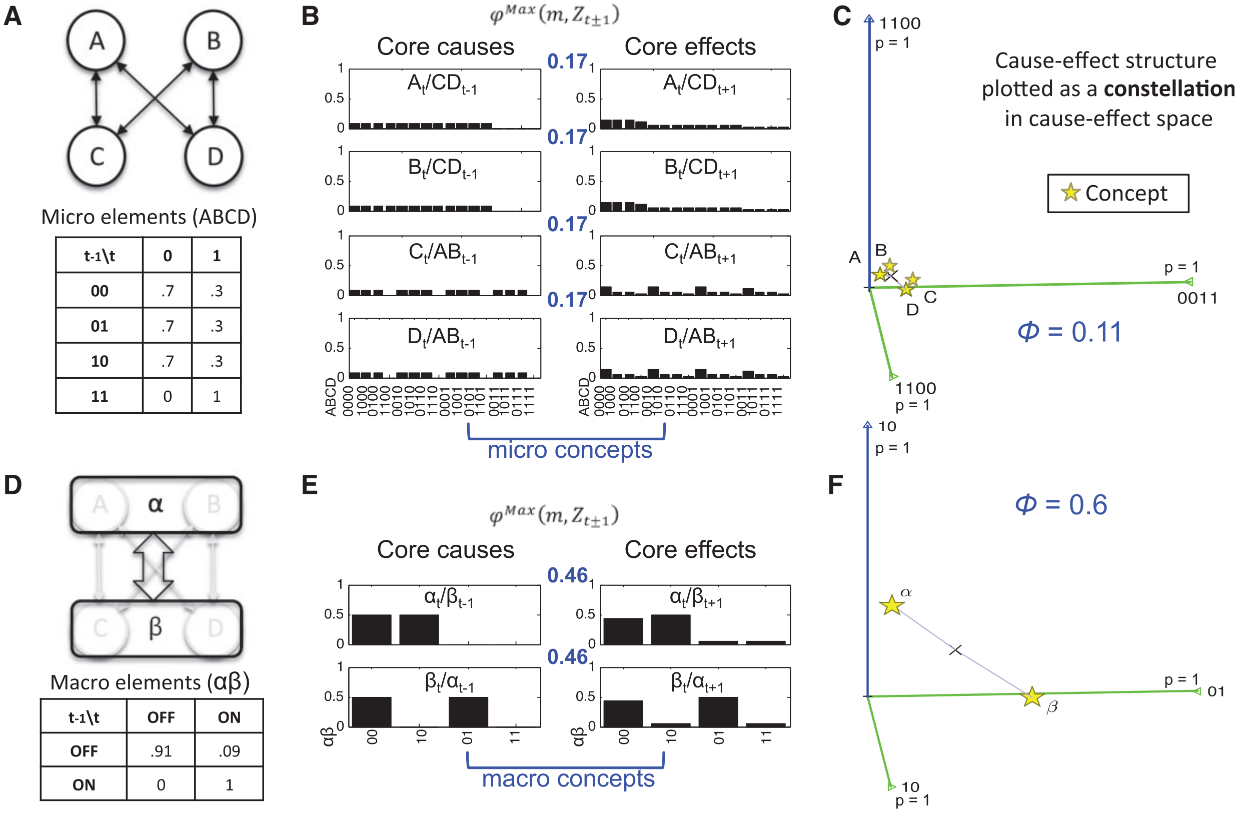
In contrast, consider the macro-level $S_M = \{\alpha, \beta\}$, shown in Fig. 2D. The micro-to-macro element mapping is of {AB} to {α} and {CD} to {β}, while the state mapping for each macro element is such that the micro states [00, 01, 10] are considered [OFF] and [11] is considered [ON]. This mapping creates the macro element tables seen at the bottom of Fig. 2D. The macro elements are each associated with a macro concept, for which $\varphi = 0.46$ (Fig. 2E). The conceptual structure of S_M (the maximally irreducible cause–effect structure of S) is plotted in cause–effect space (Fig. 2F). It shows that the two macro-level concepts are more irreducible (larger stars) and less clustered than those in S_m. The average distance (taking pairwise EMDs between the cause–effect repertoires) between all the macro concepts = 1.91, while the average distance between all the micro concepts = 0.9. $S_M = \{\alpha, \beta\}$ is the optimal binary coarse-graining with $\Phi^{\mathrm{Max}}(S_M) = 0.6$; the MIP noises the connections from micro elements {AC} to {BD}. Accordingly, Φ does not peak at the micro level, but at the macro-level $S_M = \{\alpha, \beta\}$.

How does the macro beat the micro?

Broadly, the macro beats the micro by grouping together redundant or noisy elements to increase their cause–effect power, as measured by φ. As shown in the "Theory" section, φ can be decomposed into repertoire size, irreducible selectivity, and selectivity-shift. Here, we show that while the size of the repertoire always decreases with coarse-graining, both irreducible selectivity and selectivity-shift can increase to a degree that outweighs the loss in size, which allows the macro to beat the micro. Figure 3 shows an example of the unpartitioned and partitioned cause–effect repertoires of {A}, as well as of its supervening macro concept {α}. The size of the micro concept's repertoire is 2, while the size of the macro concept's repertoire is 1. However, at the micro level, both the unpartitioned and partitioned distributions are very close to maximum entropy (shown in blue) and thus the irreducible selectivity of the micro concept is only 0.09. By comparison, the irreducible selectivity of the macro concept is 0.37. The selectivity-shift changes from 0 at the micro level to 0.09 at the macro level. Thus, while the macro loses 0.09 in φ from the loss in repertoire size, it gains 0.28 in φ from the increase in irreducible selectivity and an additional 0.09 in φ from the increase in selectivity-shift. Note that most of the gain stems from an increase in irreducible selectivity, rather than selectivity-shift. This gain in irreducible selectivity is due to an increase in determinism (because for the macro concept $\varphi = \varphi_{\mathit{Effect}}$, so irreducible selectivity is determined by the effect-repertoire). This is in line with previous results from Hoel *et al.* (2013), showing that macro-level causal relationships could have greater determinism and less degeneracy than their underlying micro-level causal relationships. As concepts with higher φ generally contribute more to Φ, systems with higher sums of φ generally have higher levels of Φ (Albantakis *et al.*, 2014; Albantakis and Tononi, 2015). Thus, the greater φ of the macro concepts allows the macro to beat the micro in terms of Φ as well.

Macro causal emergence in a highly degenerate system

Macro concepts can also have higher φ than their underlying micro concepts by having less degeneracy. Consider the deterministic micro level of the system shown in Fig. 4A. The S_m elements {A–F} are a cycle of AND gates in state [000000]. As AND gates in state [0] are highly degenerate (see Fig. 4A, table), each concept has $\varphi = 0.167$, despite having a repertoire size of 2. The low irreducible selectivity of the concepts (0.083) reflects this degeneracy. The selectivity shift for all micro concepts is 0. In cause–effect space, the micro concepts are clustered (Fig. 4B), and the MIP noises the connections from {A} to {BCDEF}, resulting in a value of $\Phi(S_m) = 0.19$. The winning macro set is $S_M = \{\alpha\beta\gamma\}$, where the pairs of {AB}, {CD}, {EF} are grouped into {α}, {β}, {γ}, respectively. For each macro element, the micro states [00, 01, 10] are considered [OFF] and [11] is considered [ON]. Thus, the cycle of AND gates at the micro level has been turned into a cycle of COPY gates at the macro level (Fig. 4C). At the macro level, the MIP noises the connections from {ABCDE} to {F}, resulting in $\Phi^{\mathrm{Max}}(S_M) = 0.83$ (Fig. 4D). For all macro concepts,



t-1\t	0	1
00	.7	.3
01	.7	.3
10	.7	.3
11	0	1

t-1\t	OFF	ON
OFF	.91	.09
ON	0	1

Figure 2. Spatial causal emergence of integrated information (increasing determinism). (A) The micro-level S_m is constituted of noisy elements. (B) In state [0000], the four micro concepts all share the same φ value. Shown are the core causes and effects; the format A_t/CD_{t-1} indicates that the concept belongs to A in its current state (c) and has a purview of CD in their past states. (C) A 3D projection of the 32D cause–effect space. The one past (blue) and two future (green) dimensions chosen were those with the greatest variance of probabilities (so the visualization maximizes the distances between concepts). The cause–effect structure of S_m appears as a clustered constellation of small (low φ) concepts. (D) The elements at the macro level of the system S_M are less noisy than those in S_m. (E) The two macro elements each generate a concept. (F). The conceptual structure of the macro-level system, plotted as a constellation in S_M's 8D cause–effect space, has larger stars (high φ), indicating greater irreducibility.

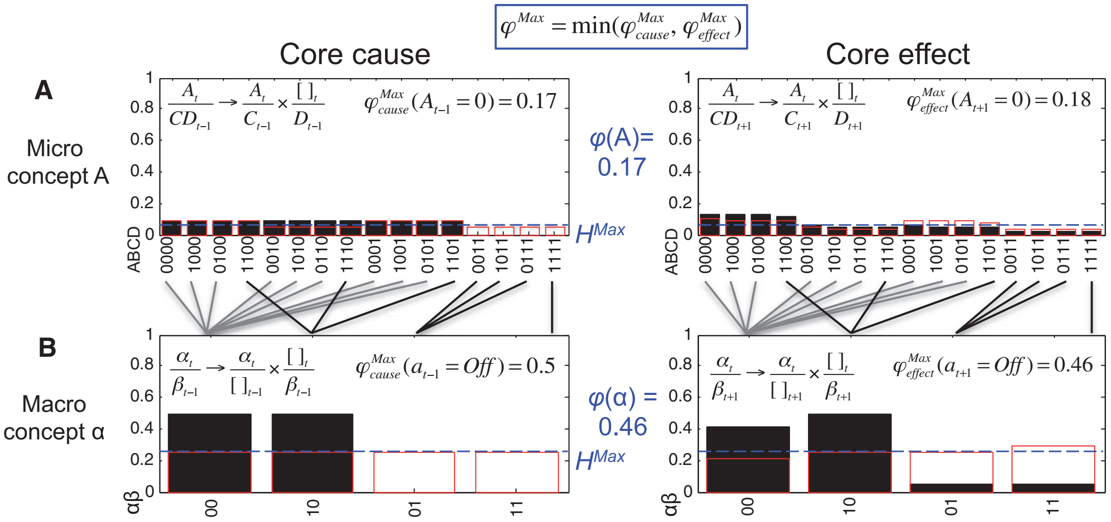
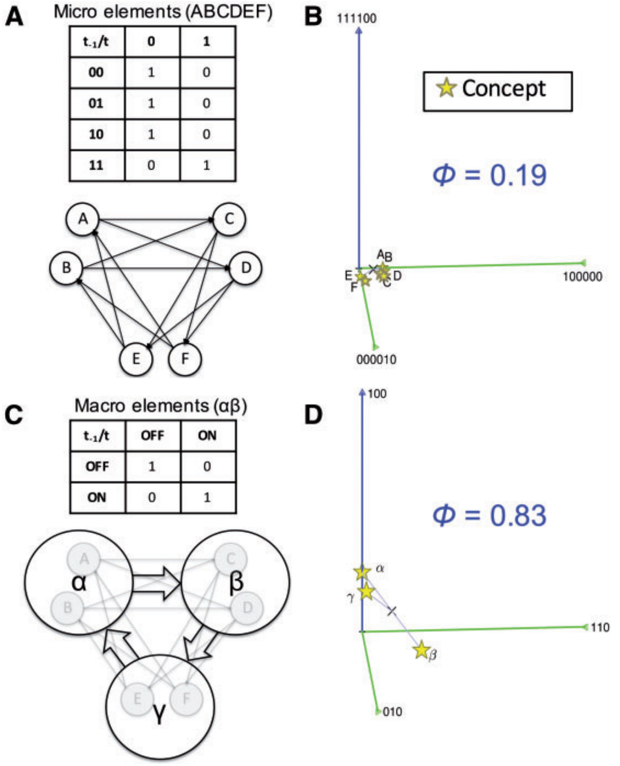


Figure 3. How the macro beats the micro. A comparison of a micro concept of Fig. 2 to its supervening macro concept. (A) The core cause–effect repertoires of element {A} which have been expanded over the whole system, with their respective MIP shown in the upper left corners. The unpartitioned repertoires are in solid black, while partitioned repertoires are in red. The dotted blue line shows where the maximum entropy distribution lies. (B) The expanded core cause–effect repertoires of element {α}, which supervenes on {A, B}. Comparing the macro cause–effect repertoire to the micro, it is obvious that selectivity (distance of the distribution to maximum entropy) is much higher for the coarse-grained mechanism, which ultimately leads to higher φ for the macro concept.



t_{-1}/t	0	1
00	1	0
01	1	0
10	1	0
11	0	1

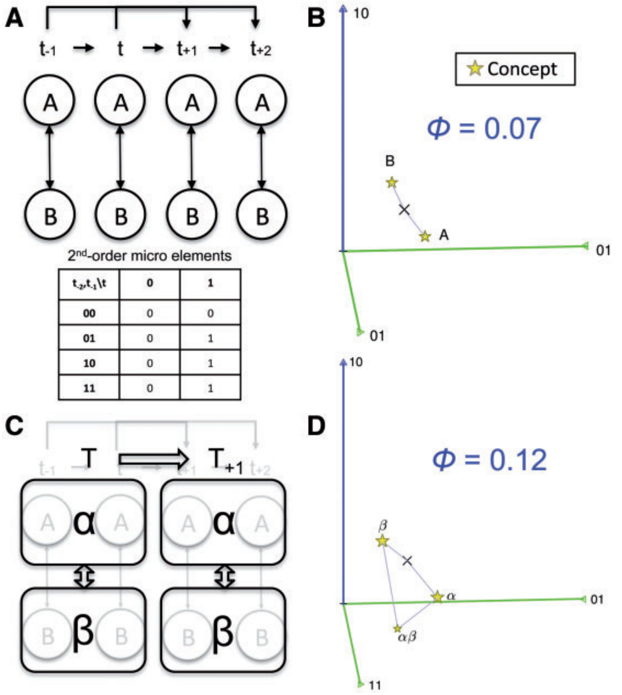
t_{-1}/t	OFF	ON
OFF	1	0
ON	0	1

Figure 4. Spatial causal emergence through degeneracy. (A) A highly degenerate but deterministic S_m is constituted of AND gates (analyzed in state [000000]). (B) The micro-level cause–effect structure is highly clustered and the concepts' irreducibility is low. (C) S_M is still deterministic but is no longer degenerate. (D) The conceptual structure is less clustered and more irreducible at the macro level.

$\varphi = 0.5$. The average distance between the macro concepts is 2, while the average distance between all the micro concepts is 1.33. While the size of the macro concepts' repertoires is reduced to 1, their irreducible selectivity has increased to 0.5 (their selectivity-shift = 0), meaning that the macro elements have 0 degeneracy. It is this reduction in degeneracy that allows the macro to beat the micro even though the system is completely deterministic.

Temporal coarse-graining and macro causal emergence

Macro groupings may be over time as well as space (Hoel *et al.*, 2013). In the temporal case, it is the micro timesteps (*t*) that are coarse-grained into macro timesteps (*T*). All possible macro timesteps are considered. The timestep over which the intrinsic cause–effect power is maximal (Φ^{Max}) is the timestep at which a system operates causally from its intrinsic perspective. For example, consider the system in Fig. 5A, {AB}, constituted of second-order Markov elements. Analysis of the system at the micro timestep (S_t), in state AB = [11], results in $\Phi = 0.07$ (Fig. 5B), with a MIP from {A} to {B}. Figure 5C shows the system analyzed at a macro timestep (S_T) wherein $\{A_{t-1}, A_t\}$ are grouped together into a single macro element over two micro timesteps $\{\alpha_T\}$, and with $\{B_{t-1}, B_t\}$ grouped into $\{\beta_T\}$. In the mapping, the micro state $A_{t-1}A_t = [00]$ is considered $\alpha_T = [\text{OFF}]$, while the micro states $A_{t-1}A_t = [01, 10, 11]$ are considered $\alpha_T = [\text{ON}]$ and likewise for the states of $B_{t-1}B_t$ and β_T. The system at the macro timestep has $\Phi^{\text{Max}}(S_T) = 0.12$, which means that from the intrinsic perspective the system operates causally over that macro timestep (two micro timesteps). The macro conceptual structure is shown in Fig. 5D. The macro MIP is $\{A_{t-1}, B_{t-1}\}$ to $\{A_t, B_t\}$. Because the micro elements are second-order Markov, unlike in the previous examples, the system at the macro timestep does not supervene on the system at a single micro timestep. This is why the repertoire size at the macro timestep (average = 2.5) is not smaller than the repertoire size at the micro timestep (average = 1). This also explains why the micro timestep is missing one of the concepts present at the macro timestep. At the macro timestep, the average irreducible selectivity = 0.16 and average selectivity-shift = 0.03; at the micro timestep, average irreducible selectivity = −0.125 and average selectivity-shift = 0.125. The average distance between macro concepts is 0.75 and between micro concepts is 0.58.



$t_{-2},t_{-1}\backslash t$	0	1
00	0	0
01	0	1
10	0	1
11	0	1

Figure 5. Temporal causal emergence. (A) S_m is constituted of second-order Markov elements. (B) Analysis at the micro timestep, in which AB = [11], specifies a cause–effect structure with two clustered concepts with low irreducibility. (C) The macro timestep $\{\alpha,\beta\} = [\text{On, On}]$ is a grouping over two micro timesteps. (D) The macro conceptual structure, which is more irreducible than the micro conceptual structure, is less clustered and has an additional concept (see text).

Complexes are maxima of irreducible, intrinsic cause–effect power over both elements and spatiotemporal grains

A complex is the set of elements that specifies a maximum of Φ across all spatiotemporal scales and defines the elements and borders of the physical system. The set of elements with maximal Φ at a particular spatiotemporal level may differ across spatiotemporal grains. For example, consider the system in Fig. 6A in state [000000]. If the analysis is restricted to the micro level, the complex would be identified over the full set of elements {A–F}, with a total of eight micro concepts (Fig. 6B). In this analysis, Φ is only 0.15 since, despite the average size of their repertoires being 2.25, the concepts have low irreducible selectivity (0.09), demonstrate little selectivity-shift (0.02), and are not very

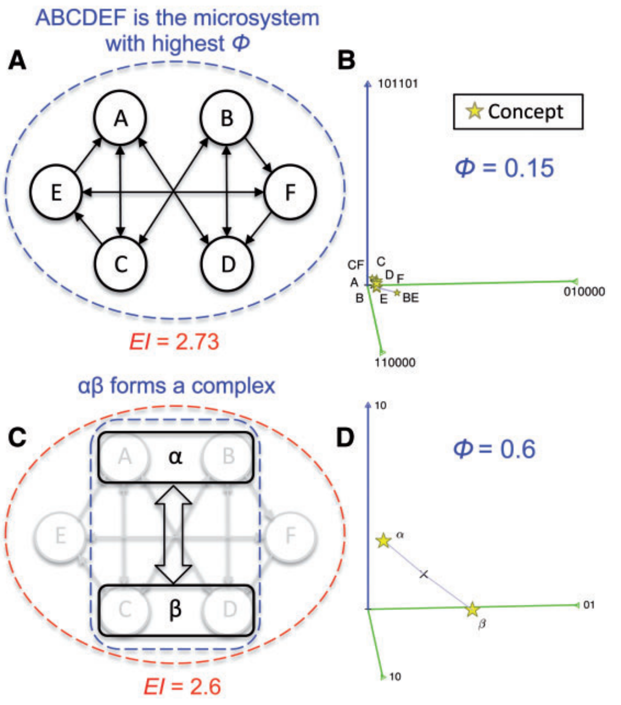


Figure 6. Macro borders versus micro borders. (A) S_m is constituted of heterogeneous logic gates: {ABCD} act identically to those described in Fig. 2, while {EF} each act as deterministic AND gates. Additionally, if {E} = 1 at t_{-1}, then the probability of {A} = 1 at t increases from 0.3 to 0.9; the same rule applies for the connection from {F} to {D}. In state [000000], the set of micro elements with highest Φ is {A–F} (dotted blue line). EI of S_m in red. (B) The maximally irreducible cause–effect structure of S_m with eight concepts. (C) The complex is at the level of $S_M = \{\alpha,\beta\}$ (dotted blue line) and only supervenes on a subset of the micro elements {ABCD}. The highest EI of all possible coarse-grainings is shown in red. Note that EI is always over the whole system (dotted red line). (D) The macro conceptual structure has $\Phi^{Max} = 0.6$.

distant in cause–effect space (average = 1.27). The MIP is from {EF} to {ABCD}. However, extending the search for Φ^{Max} across all candidate sets at all possible coarse-grainings finds the true complex consists of the macro elements {αβ} with $\Phi^{Max}(S_M) = 0.6$. Note that this is a grouping of micro elements {ABCD} that does not include {EF}. The state grouping is the same as in Fig. 2, with the same macro repertoire size (1), irreducible-selectivity (0.37), selectivity-shift (0.09), and average distance (1.91). According to IIT, only the maximum of Φ over both elements and levels qualifies as a complex (other elements and levels are excluded). That is, Φ^{Max} represents the absolute maximum over all values of Φ of all possible groupings found at each spatial and temporal scale. Therefore, from the intrinsic perspective a system "self-defines," independent of an external observer, both its borders (the set of elements that are included in the complex) and its spatiotemporal level.

The example in Fig. 6 highlights that evaluating macro causal emergence by assessing cause–effect power in full, using a measure of integrated information Φ that takes into account composition, integration, exclusion, and state-dependency, is more accurate than using EI (Hoel *et al.*, 2013). For this system, EI fails to show causal emergence ($EI(S_m) = 2.73 > EI(S_M) = 2.6$). One reason for this is that EI cannot identify causal borders (exclusion), so EI is always over the whole system (ABCDEF or a coarse-graining thereof). Another reason is that, while EI is a state-independent measure, Φ is state-dependent: for example, in state [000000] the system demonstrates macro causal emergence as shown above, but in state [111111] the micro level has Φ^{Max} ($\Phi^{Max}(S_m) = 1$, $\Phi(S_M) = 0.21$).

Trends of integrated information across spatiotemporal grains point to the level at which maxima occur

In Fig. 7, we show Φ values at each possible binary spatial and temporal level for the four example systems. The winning macro levels appear as clear maxima for all of the systems. We next examined the relationships between each level of the system. That is, we tracked how higher level macros are coarse-grains of coarse-grains. For example, consider a micro level {ABCD}, for which the micro elements {AB} are grouped into a macro element {α} with a state mapping of [00, 01, 10] into [OFF] and [11] into [ON]. A further coarse-grain might group {CD} into {β} with the same state grouping, and then the next coarse-grain might group {αβ} together into a single macro element. This nesting of coarse-grains reveals paths of coarse-graining that span all the spatiotemporal levels, from the finest to the coarsest. Intriguingly, the maximal Φ value at each level of coarse graining almost always lies on the path from the micro level to the coarse-grain with Φ^{Max}. Additionally, instead of being distributed randomly, Φ increases closer to the level at which Φ^{Max} occurs. These observations suggest that, in large systems, it may be useful to assess integrated information by gradient ascent to more rapidly converge onto the optimal spatiotemporal grain.

Discussion

In this article, we applied the mathematical framework of IIT (Oizumi *et al.*, 2014) to simple example systems at multiple levels of coarse-graining across elements and timesteps. We have presented several examples for which integrated information (Φ) is highest at a macro scale, with spatial and/or temporal coarse-grains, in systems that are either deterministic or stochastic. In this way, we have provided a proof of principle that, under certain conditions discussed below, the macro can beat the micro in terms of Φ, a measure of intrinsic cause–effect power, which takes into account the causes and effects within a system that matter the most for the system itself. In line with Hoel *et al.* (2013), our results thus show that the reductionist assumption of greatest cause–effect power at the micro level is often unjustified. Moreover, our findings open up the possibility, advocated by IIT, that there is a macro spatiotemporal scale at which neural interactions within the brain have maximal intrinsic cause–effect power, corresponding to the macro spatiotemporal scale of the neural substrate of consciousness.

How the macro can beat the micro

How can the macro be causally more powerful than the micro, even though macro elements and intervals are constituted of micro elements and intervals, and even though macro properties supervene upon micro properties—that is, no extra macro causal property is introduced?

It has been suggested before (Yablo, 1992; List and Menzies, 2009) that a higher-level description of a cause may be a better account in cases where multiple lower-level conditions (counterfactuals) could all count as sufficient causes. Under a notion of causation as difference-making, it may then be appropriate

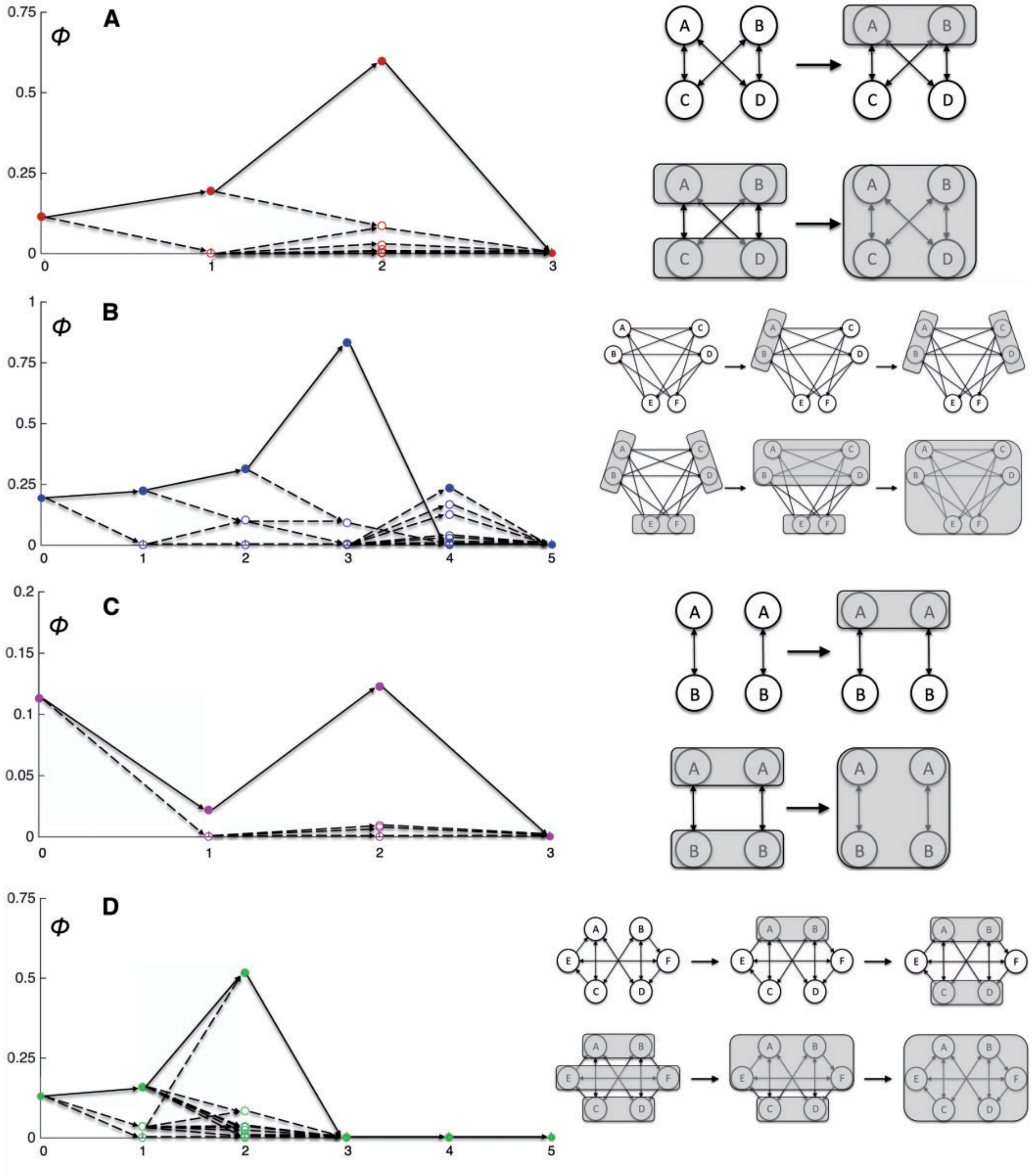


Figure 7. Finding spatial and temporal maxima. (A) All possible spatial groupings of the four-element system from Fig. 2. (B) All possible spatial groupings of the six-element system from Fig. 4. (C) All possible temporal groupings for the two-element system in Fig. 5. (D) All possible spatial groupings for the six-element system in Fig. 6. For each of the four systems previously examined, the coarse grains at each possible spatiotemporal level are plotted against their Φ values (y-axes) on the left of the figure. The x-axes represent the levels of the system: level 0 being the micro level, level 1 being a single grouping of two micro elements and so on as the degree of coarse-graining increases until all the elements have been grouped into one macro element (which always has a $\Phi = 0$). In the left plots, the solid color data points represent the maximum Φ value of all the groupings at that particular level. The relationships between levels are shown as arrows: each represents a further grouping of a lower level. Note that in each example system the path from the micro level to the grouping with Φ^{Max} (tracked by solid arrows) includes the maximum Φ value at almost every level (except the example shown in B at level 4).

to reverse the reductionist exclusion principle, allowing higher-level causes to exclude the lower-level ones (List and Menzies, 2009). In a recent study (Hoel *et al.*, 2013), we provided the first quantitative demonstration of this idea using "effective information (EI)", a measure that rigorously quantifies differences that make a difference within a system under all possible system perturbations (all counterfactuals). In simple examples of physical systems, we assessed EI at the micro and all possible coarse-grained macro levels, and could show that EI among a system's elements can indeed be greatest at a macro rather than the micro level—a form of true causal emergence of the macro over the micro. Despite the necessarily smaller state space of the coarse-grained macro level, “the macro can beat the micro” if the micro level has greater indeterminism and/or degeneracy than the macro level (Hoel *et al.*, 2013).

Integrated information (Φ) assesses several key features of cause–effect power that are not captured by EI (Oizumi *et al.*, 2014): the dependence of cause–effect power on the specific state the system is in (state-dependency); the cause–effect power of the system’s parts (composition); whether the whole system is causally irreducible to its parts (integration); and what defines the system’s causal borders (exclusion). In measuring Φ, the first step is to evaluate the cause–effect repertoires specified by all subsets of elements within a system. Each subset of elements in a state (a “mechanism”) constrains past and future states within its candidate system, and the extent of this constraint is the mechanism’s selectivity. The integrated information of a mechanism (φ), in turn, measures to what extent a cause–effect repertoire specified by the mechanism is irreducible to that specified by its parts. Maximally irreducible cause–effect repertoires are called concepts (see “Theory” section). The present analysis distinguishes three features of the cause–effect repertoire of a concept—repertoire size, irreducible selectivity, and selectivity-shift. Repertoire size is a function of the number of elements whose past and future states are constrained by a concept. Since there are necessarily fewer macro than micro elements (or intervals), macro-level concepts have an a priori lower capacity for φ. Coarse-grains are thus at a disadvantage in terms of cause–effect power. However, in line with previous results (Hoel *et al.*, 2013), we show here that macro mechanisms can make up for their reduced repertoire size by achieving higher selectivity.

As we moreover confirmed, this result holds even though integrated information postulates the additional requirement of irreducibility (integration): Macro-level mechanisms can achieve greater irreducible selectivity. This means that the macro can win if macro mechanisms constrain past and future macro states irreducibly—above and beyond their parts—to a far greater extent than micro mechanisms do. Selectivity-shift further accounts for any change in the probability of past and future states irrespective of changes in irreducible selectivity, though its contribution in the examples presented in this article is minor. Finally, the example of Fig. 5 shows that the winning macro system can contain irreducible higher-order mechanisms (composed of two elements) that have no counterpart in the best micro system.

A system’s spatiotemporal scale from its intrinsic perspective

As mentioned in the Introduction, integrated information captures several essential features of cause–effect power that go above and beyond EI, namely state-dependency, composition, integration, and exclusion. These features make Φ a measure of “intrinsic” cause–effect power, which according to IIT is a requirement for identifying the PSC of consciousness (Tononi, 2012, 2015; Oizumi *et al.*, 2014). As previously shown, these aspects are also essential for understanding how the causal structures of simulated organisms evolve in a simulated environment (Albantakis *et al.*, 2014) and for the causal analysis and classification of discrete dynamical systems, such as cellular automata (Albantakis and Tononi, 2015). Here, we show their importance for establishing conclusively if and when macro beats micro.

Unlike EI, Φ assesses the combinatorial contribution of the system’s parts on its cause–effect power (composition), which is significant in assessing coarse-grains given that a system has vastly more parts, hence combinations, at the micro level. Crucially, evaluating irreducibility at every spatiotemporal grain ensures that we do not attribute cause–effect power to collections of elements that causally are nothing more than the sum of their parts (integration). Moreover, assessing maxima of irreducible cause–effect power ensures that we identify the borders that define a causal system at each spatiotemporal grain (exclusion). This is equivalent to requiring that the cause–effect power of each element is counted only once (Oizumi *et al.*, 2014). The search for a maximum of integrated information (Φ^{Max}) thus identifies a definite spatiotemporal scale at which a set of elements “self-defines” as a complex—the grain size at which it “comes into focus” causally from its own intrinsic perspective. Such a grain size is determined by the intrinsic cause–effect structure of the system itself, as opposed to being the most convenient or interesting scale for an external observer.

Since integrated information Φ evaluates not just average cause–effect power, but additionally takes into account state-dependency, composition, integration, and exclusion, it is a much more sensitive measure of intrinsic cause–effect power than EI. For example, an analysis purely in terms of EI (Hoel *et al.*, 2013) of the system in Fig. 6 would have concluded that the overall system has maximum cause–effect power at the micro level. Instead, the present analysis shows that the system is actually constituted of a macro-level complex (α,β) that supervenes on (ABCD), but does not include the micro-level elements (EF). As determined by Φ^{Max}, the macro complex (α,β) is a system from its own intrinsic perspective.

Assessing the intrinsic spatiotemporal scale of the brain: limitations and heuristic strategies

An open question in neuroscience is whether there is a particular spatial and temporal grain at which the brain “works.” Consider the many cognitive functions carried out by the cerebral cortex: does every neuron matter, or only groups of neurons? Does every spike count, or only synchronous activity over tens of milliseconds? From the extrinsic perspective of an observer, the spatiotemporal scale of measurement drastically affects whether an event is information rich or scarce (Panzeri *et al.*, 2010). The approach presented here suggests a principled way of determining how the brain decomposes into causal subsystems and at which respective spatiotemporal grains from its own intrinsic perspective. In other words, while a neurophysiologist can investigate the cerebral cortex at every level, from quantal release of transmitters to the functional connectivity among entire brain regions, from the intrinsic perspective there is a privileged spatiotemporal grain at which cause–effect power is exerted.

Given practical and computational constraints, using Φ to evaluate the intrinsic spatiotemporal scale of the brain across all its possible levels and sets of elements is infeasible. All the examples discussed in this study are small, idealized binary systems for which cause–effect structures and associated Φ values could be assessed rigorously at all spatiotemporal grains. Even for such simple, completely characterized systems, calculating Φ across all levels is computationally demanding. For larger systems, and certainly for real systems that are not completely characterized at the microphysical level, an exhaustive assessment of Φ is problematic. Nevertheless, heuristic criteria can be employed to identify a range of coarse-grainings with high likelihood of yielding high Φ values. In the brain, this range likely includes individual neurons or groups of neurons in space, and milliseconds to a few seconds in time, as it is unlikely that finer grainings would yield substantial integration or that coarser grainings would yield sufficient selectivity. Within a population of neurons, assessing whether intrinsic cause–effect power is higher at the level of individual neurons or groups of neurons would be difficult but not impossible given currently available neurophysiological and optogenetic tools. Assuming a coarse-grained state space, for example, in which each neuron has three relevant states ("silent," "firing," and "bursting"), the TPM of a population of neurons could be obtained by using optogenetics to perturb elements and calcium imaging to observe their transitions. From this TPM intrinsic cause–effect power (Φ) can, in principle, be evaluated across all levels of coarse-graining (see "Theory" section).

If applying the full IIT computation were too prohibitive, one could still assign neurons into distinct macro-groups guided by heuristic criteria (e.g. similar receptive fields, etc.) and compare the irreducible selectivity of their macro cause–effect repertoires against those of individual neurons. The present results indicate that, in every example considered, there is a "path" of coarse-grains such that tracing the highest Φ values at each level eventually leads to the overall maximum value of Φ (Fig. 7). This suggests that, by performing iterative searches by gradient ascent, it may be possible to more rapidly converge onto the optimal spatiotemporal grain, which would allow assessing a greater number of potentially relevant coarse-grainings in the brain. This would indicate whether the macro level could indeed make more of a difference from the intrinsic perspective of the neuronal population itself.

At the more coarse-grained levels of smaller or larger brain regions, neuroimaging studies can be used to evaluate two key requisites for high Φ: the informational capacity of a system can be approximated by measures that capture neurophysiological "differentiation," the repertoire of distinct neural states the system can visit (Boly *et al.*, 2015; Marshall *et al.*, 2016); integration can be assessed through measures of functional or effective connectivity between brain regions (Seth et al., 2011; Boly *et al.*, 2012). Neurophysiologically realistic, large-scale computer simulations (Deco *et al.*, 2014) are another, complementary approach to assess intrinsic cause–effect power as they enable investigating the effects of perturbations at various levels of coarse-graining.

The fact that the current framework can only be applied to discrete systems is another limitation. However, even if truly continuous systems were to exist, any causal analysis based on perturbation and measurement would be necessarily coarse-grained, as it is impossible to perturb a continuous system into every state with equal likelihood. Even so, it may still be possible to show that Φ values decrease when fine-graining perturbation and measurement toward their physical limit.

Finally, in the present work, each macro element was defined by coarse-graining micro elements, i.e. by averaging over all their inputs and outputs (states). This approach may be adequate when considering, for example, whether within some parts of the brain neurons with similar response properties, inputs, and outputs have more cause–effect power taken as groups or as individual neurons. Coarse-graining is not appropriate, however, for determining if a neuron, taken as a macro element, has more cause–effect power than the set of specifically organized molecules (or even smaller micro elements) that constitute it. To evaluate macro causal emergence in such cases, instead of coarse-graining, one needs to "black box" many micro elements within a macro element and consider just a few inputs and outputs to/from the black box (e.g. the neuron). Furthermore, the present work focuses on the role of selectivity in understanding how the macro can beat the micro, by strengthening the cause–effect power of mechanisms in the system. Another possible way for the macro to beat the micro is by specifying additional mechanisms (especially higher-order mechanisms) at the macro level that are not specified at the micro level, a possibility not explored in the current work. These topics will be the subject of a future publication (Marshall *et al.*, 2016).

The spatiotemporal grain of consciousness

Establishing the spatiotemporal grain at which intrinsic cause–effect power peaks in the brain is not only important in its own right but, according to IIT, it is directly relevant for characterizing the neural substrate of consciousness. In fact, measures of integrated information were originally developed with the explicit purpose of characterizing the requirements for physical systems to be conscious (Tononi, 2004, 2008, 2012; Oizumi *et al.*, 2014). Specifically, based on phenomenological axioms, IIT claims that the PSC of consciousness is a set of elements in a state, at a particular spatiotemporal grain, that specifies a cause–effect structure having maximally irreducible, compositional, intrinsic cause–effect power (Tononi, 2012; Tononi *et al.*, 2016). Currently, Φ-related measures (Barrett and Seth, 2011; Oizumi *et al.*, 2016) and other indices of causal effectiveness are already being applied in both theoretical and empirical studies of consciousness (Seth, 2008; Seth et al., 2011; Casali *et al.*, 2013). For this purpose, it is important to establish the spatiotemporal grain of the neural elements constituting the neural substrate of consciousness and thereby account for why consciousness occurs at the particular spatiotemporal scale it does (Tononi, 2004; Marom, 2010; Chalmers, 2013).

It is not currently known whether neurons or groups of neurons at coarser or finer grains of activity form the units corresponding to phenomenological distinctions (Tononi *et al.*, 2016). In principle, using an approach similar to the one presented here, it should be possible to assess at which spatiotemporal grain integrated information reaches a maximum in the brain. With experimental techniques like those outlined above, one could now test IIT's prediction that the neurophysiological maximum of cause–effect power corresponds to the spatiotemporal scale of experience (Bachmann, 2000; Holcombe, 2009). If the experimental evidence indicates that neuronal groups rather than single neurons constitute the scale of maximal intrinsic cause–effect power in the brain, IIT would predict that changes in the average activity of a group of neurons should make a difference to the content of experience, while changes to individual neurons that do not affect the average group activity should not (Tononi *et al.*, 2016).

Another interesting question that could be addressed is to what extend the spatiotemporal grain and relevant activity states of the elements with maximal integrated information (Φ^{Max}) vary across brain regions, time, different states of consciousness (wake, dreamless sleep, anesthesia), and even different tasks or attentional states.

Finally, if the prediction were validated by studies in humans, one could extrapolate to the spatiotemporal scale of experience in other species, at least some of which are known to integrate sensory signals at a temporal scale that is different from ours (Healy *et al.*, 2013).

Highlights

- Integrated information can be measured in systems at different spatial and temporal scales.
- Integrated information is a state-dependent measure of causal power from the intrinsic perspective of the system.
- This approach provides a way to assess the spatiotemporal scale of the physical substrate of consciousness.

Supplementary data

Supplementary data is available at *Neuroscience of Consciousness Journal* online.

Acknowledgments

We thank Will Mayner for his work on the Python IIT software (PyPhi). This work has been supported by the Templeton World Charity Foundation (grant no. TWCF 0067/AB41).

Conflicts of interest statement. The authors declare no conflicts of interest.

References

Albantakis L, Hintze A, Koch C *et al*. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Comput Biol* 2014;**10**:e1003966.

Albantakis L, Tononi G. The intrinsic cause-effect power of discrete dynamical systems—from elementary cellular automata to adapting animats. *Entropy* 2015;**17**:5472–502.

Bachmann T. *Microgenetic Approach to the Conscious Mind*. Amsterdam: John Benjamins Publishing, 2000.

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

Boly M, Sasai S, Gosseries O *et al*. Stimulus set meaningfulness and neurophysiological differentiation: a functional magnetic resonance imaging study. *PLoS One* 2015;**10**:e0125337.

Boly M, Massimini M, Garrido MI *et al*. Brain connectivity in disorders of consciousness. *Brain Connect* 2012;**2**:1–10.

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.

Chalmers D. The combination problem for panpsychism. In: Bruntrup G. and Jaskolla L. (ed.) *Panspsychism Reef*. Oxford University Press, 2013.

Deco G, Hagmann P, Hudetz AG *et al*. Modeling resting-state functional networks when the cortex falls asleep: local and global changes. *Cereb Cortex* 2014;**24**:3180–94.

Healy K, McNally L, Ruxton GD *et al*. Metabolic rate and body size are linked with perception of temporal information. *Anim Behav* 2013;**86**:685–96.

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

Holcombe AO. Seeing slow and seeing fast: two limits on perception. *Trends Cogn Sci* 2009;**13**:216–21.

Kim J. *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press, 1993.

Kim J. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press, 2000.

List C, Menzies P. Nonreductive physicalism and the limits of the exclusion principle. *J Philos* 2009;**106**:475–502.

Markram H. The blue brain project. *Nat Rev Neurosci* 2006;**7**:153–60.

Marom S. Neural timescales or lack thereof. *Prog Neurobiol* 2010;**90**:16–28.

Marshall W, Ramirez-Gomez J, Tononi G. Integrated information and state differentiation. *Front Psychol* 2016;**7**:926.

Marshall W, Albantakis L, Tononi G. Black boxing and cause-effect power. 2016. arXiv:1608.03461 [q-bio.NC]

Mayner W, Marshall W. Pyphi: Hotfix 0.7.1. doi:10.5281/zenodo.21149. 2015.

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.

Oizumi M, Amari S, Yanagawa T *et al*. Measuring integrated information from the decoding perspective. *PLoS Comput Biol* 2016;**12**:e1004654.

Panzeri S, Brunel N, Logothetis NK *et al*. Sensory neural codes using multiplexed temporal scales. *Trends Neurosci* 2010;**33**:111–20.

Pele O, Werman M. Fast and robust earth mover's distances. In: *Proceedings of IEEE International Conference on Computer Vision*. Kyoto: IEEE Computer Society, 2009. doi: 10.1109/ICCV.2009.5459199.

Pearl J. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press, 2000.

Rubner Y, Tomasi C, Guibas LJ. Earth mover's distance as a metric for image retrieval. *Int J Comput Vis* 2000;**40**:99–121.

Seth A. Measuring emergence via nonlinear Granger causality. *alife* 2008;545–52.

Seth AK, Barrett AB, Barnett L. Causal density and integrated information as measures of conscious level. *Philos Trans a Math Phys Eng Sci* 2011;**369**:3748–67.

Sporns O, Tononi G, Kötter R. The human connectome: A structural description of the human brain. *PLoS Comput Biol* 2005;**1**:e42.

Stalnaker R. Varieties of supervenience. *Philos Perspect* 1996;**10**: 221–41.

Tononi G, Sporns O, Edelman GM. Measures of degeneracy and redundancy in biological networks. *Proc Natl Acad Sci* 1999;**96**: 3257–62.

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

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

Tononi G. Integrated information theory of consciousness: An updated account. *Arch Ital Biol* 2012;**150**:56–90.

Tononi G. Integrated information theory. *Scholarpedia* 2015;**10**: 4164.

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

Yablo S. Mental causation. *Philos Rev* 1992;**101**:245–80.