entropy

MDPI

Article

Scaling Behaviour and Critical Phase Transitions in Integrated Information Theory

Miguel Aguilera [1,2]

[1] IAS-Research Center for Life, Mind and Society, University of the Basque Country, 20018 Donostia, Spain; sci@maguilera.net

[2] ISAAC Lab, Aragón Institute of Engineering Research, University of Zaragoza, 50018 Zaragoza, Spain

Received: 14 March 2019; Accepted: 30 November 2019; Published: 5 December 2019

Abstract: Integrated Information Theory proposes a measure of conscious activity (Φ), characterised as the irreducibility of a dynamical system to the sum of its components. Due to its computational cost, current versions of the theory (IIT 3.0) are difficult to apply to systems larger than a dozen units, and, in general, it is not well known how integrated information scales as systems grow larger in size. In this article, we propose to study the scaling behaviour of integrated information in a simple model of a critical phase transition: an infinite-range kinetic Ising model. In this model, we assume a homogeneous distribution of couplings to simplify the computation of integrated information. This simplified model allows us to critically review some of the design assumptions behind the measure and connect its properties with well-known phenomena in phase transitions in statistical mechanics. As a result, we point to some aspects of the mathematical definitions of IIT that 3.0 fail to capture critical phase transitions and propose a reformulation of the assumptions made by integrated information measures.

Keywords: Integrated Information Theory; Phi; Ising model; criticality; phase transitions

1. Introduction

Integrated Information Theory (IIT [1]) was developed to address the problem of consciousness by characterizing its underlying processes in a quantitative manner. It provides a measure of integration, Φ, that quantifies to what extent a dynamical system generates information that is irreducible to the sum of its parts, considered independently. Beyond IIT as a theory of consciousness, Φ has received attention as a general measure of complexity, and different versions of the measure have been applied to capture to what extent the behaviour of a system is both differentiated (displaying diverse local patterns) and integrated (maintaining a global coherence). Furthermore, Φ attempts to capture the level of irreducibility of the causal structures of a system, revealing the boundaries in the organisation of complex dynamical systems (i.e., delimiting the parts of the system that are integrated into a functional unit [2]).

Despite its promising features, IIT is still controversial and it has received different critiques, both to its theoretical and philosophical foundations (e.g., see [3,4]) and the mathematical definitions therein. Some of the latter suggest that Φ might not be well defined and presents a number of problems [5]. Furthermore, there is a myriad of different definitions of Φ, and experimental testing of these competing definitions in small systems shows that their properties radically diverge in some cases [6]. Despite the abundance of critiques and alternative definitions of Φ, it is not clear which is the appropriate direction to settle theoretical differences or test different approaches experimentally. In our view, there is two main obstacles that hinder this endeavour: (1) the difficulty of testing integration measures in well-known dynamical systems where integrated information measures can be evaluated

against known properties of the system, and (2) the computational cost of calculating IIT measures, preventing its application beyond small networks.

The first problem is, in general, difficult, as even in simple nonlinear models the relation between the network topology and dynamics is complex, and it is not always clear which topology should yield larger integrated information. Nevertheless, there is a family of models in which the relation between segregation and integration is well characterised: homogeneous systems exhibiting order-disorder critical phase transitions [7]. In these systems, there is a transition in their phase space from a disordered state, in which activity is random and segregated, to an ordered one, where units of the system are strongly coordinated. Just at the boundary separating ordered and disordered dynamics we find criticality, a state where a compromise between coordination and segregation is found [7,8]. Even in simple systems, critical dynamics are dominated by small bursts of local (i.e., segregated) activity, yet display large avalanches of globally coordinated (i.e., integrated) activity. In neural systems, these are generally referred to as "neuronal avalanches", and experimental evidence suggests that neural dynamics exhibit a degree of concordance with those expected for a system near criticality [9].

Critical phenomena are theoretically characterised for some systems of infinite size, and they can be experimentally identified as divergent tendencies in large, finite systems as they grow in size. This refers us to the second problem, which is the very large computational cost of measuring integrated information. Due to combinatorial explosion, computing Φ is only possible in general for small, discrete systems. In practice, this prevents to measure integrated information in the very large or even infinite systems where critical dynamics can be appreciated. In IIT 3.0 [1], the latest version of the framework of Integrated Information Theory, Φ can only be computed for binary networks composed of up to a dozen units. There is a variety of other measures to compute integrated information [6], and some of them are computationally lighter, but all share these limits to some extent. As most versions of Φ require computing distance measures between probability distributions of a system and finding the minimum information partition (MIP), they present important restrictions in terms of computational cost as a system scales in size. Note that some simplified measures of Φ can be computed analytically for the case of Gaussian distributions (e.g., see [10,11]), but, in this paper, we will focus in the general case of binary variables, without considering the specific case of Gaussian assumptions.

In general, IIT measures have been limited to very small systems, and it is not well understood how integrated information scales with the size or temporal scale of the system. Previous work [12] has analysed how integrated information changes with spatial and temporal coarse graining in small networks, but it is still difficult to connect these results with well-known scaling effects like the ones that take place in critical phase transitions. Luckily, a family of simplified models, generally referred to as Ising models, can capture critical transitions with very simple topologies. Some of the simplest ones present homogeneous architectures that can greatly simplify the calculation of information theoretical quantities of the system, see, e.g., [13]. Using this idea, recent novel work using mean-field approximations [14] has shown that under some assumptions it is possible to compute integrated information in infinite-size networks with some homogeneous properties, showing that integrated information measures diverge at certain critical points. In this work, we extend those results by finding methods for computing integrated information of similar models of finite size. Specifically, we explore how integrated information measures scale with size in large networks, proposing methods for simplifying the computation of distance metrics between probability distributions and the search of the MIP. In doing so, we critically assess some of the assumptions behind IIT measures and propose some modifications to better capture the properties of second-order phase transitions. Specifically, we will explore different aspects of the definition of integrated information: (1) the dynamics and temporal span of integrated information, (2) assumptions for the computation of the cause repertoire, (3) the choice of distance metrics between the Wasserstein distance and the Kullback–Leibler divergence, (4) the effect of considering the environment from a situated perspective, (5) the relation between mechanism-level integration φ and system-level integration Φ and (6) the importance of identifying diverging tendencies in the system.

2. Model

To show how integrated information behaves around critical points in Ising models, we describe a (slightly modified) version of IIT 3.0. Then, we introduce a family of homogeneous kinetic Ising models and a series of methods that simplify the computation of integrated information in large networks.

2.1. IIT 3.0

We critically revise and adapt the framework of IIT 3.0 [1], which originally computes the integrated information of a subset of elements of a system as follows. For a system of N elements with state $\mathbf{s}$ at time t (We use boldface letters and symbols for vectors and matrices, e.g., $\mathbf{s}(t) = (s_1(t), s_2(t), \ldots, s_N(t))^T$, ^{T} indicating transposition.), we characterise the input–output relationship of the system elements through its corresponding transition probability function $P(\mathbf{s}(t+\tau)|\mathbf{s}(t))$, describing the probabilities of the transitions from one state to another for all possible system states. IIT 3.0 requires systems to satisfy the Markov property (i.e., that the distribution of states of a process at time t depends only upon the state at time $t-\tau$), and that the current states of elements are conditionally independent, given the past state of the system, i.e., $P(\mathbf{s}(t)|\mathbf{s}(t-\tau)) = \prod_i P(s_i(t+\tau)|\mathbf{s}(t))$.

IIT 3.0 computes integrated information in the causal mechanisms of a system by defining two subsets of $\mathbf{s}(t)$ and $\mathbf{s}(t \pm \tau)$, called the mechanism $\mathcal{M}_t = \{s_i(t)\}_{i \in I_{\mathcal{M}_t}}$ and the purview $\mathcal{P}_{t\pm\tau} = \{s_i(t \pm \tau)\}_{i \in I_{\mathcal{P}_{t\pm\tau}}}$, to represent the current state of part of the system and how it constrains future or past states. The cause-and-effect repertoires of the system are described, respectively, by the probability distributions $P(\mathcal{P}_{t-\tau}|\mathcal{M}_t)$ and $P(\mathcal{P}_{t+\tau}|\mathcal{M}_t)$.

The integrated cause–effect information of $\mathcal{M}_t$ is then defined as the distance between the cause–effect repertoires of a mechanism and the cause–effect repertoires of its minimum information partition (MIP) over the maximally irreducible purview,

$$\varphi_{\text{cause}}(\tau) = \max_{\mathcal{P}} \left(\min_{\text{cut}} \left(D_W(P(\mathcal{P}_{t-\tau}|\mathcal{M}_t), P_{\text{cut}}(\mathcal{P}_{t-\tau}|\mathcal{M}_t)) \right) \right), \tag{1}$$

$$\varphi_{\text{effect}}(\tau) = \max_{\mathcal{P}} \left(\min_{\text{cut}} \left(D_W(P(\mathcal{P}_{t+\tau}|\mathcal{M}_t), P_{\text{cut}}(\mathcal{P}_{t+\tau}|\mathcal{M}_t)) \right) \right), \tag{2}$$

where $D_W(P, Q)$ refers to the Wasserstein distance (also known as Earth Mover's Distance), used by IIT 3.0 to quantify the statistical distance between probability distributions P and Q. The subindex cut specifies a bipartition of the mechanism into two halves and P_{cut} represents the cause or effect probability distribution under such partition,

$$\begin{aligned} \text{cut} &= \left\{ \mathcal{M}_t^1, \mathcal{P}_{t\pm\tau}^1, \mathcal{M}_t^2, \mathcal{P}_{t\pm\tau}^2 \right\}, \\ P_{\text{cut}}(\mathcal{P}_{t\pm\tau}|\mathcal{M}_t) &= P(\mathcal{P}_{t\pm\tau}^1|\mathcal{M}_t^1) \otimes P(\mathcal{P}_{t\pm\tau}^2|\mathcal{M}_t^2). \end{aligned} \tag{3}$$

Here, cut specifies the partition applied over the elements of mechanism $\mathcal{M}$, where $\mathcal{M}_t^1, \mathcal{M}_t^2$ design the blocks of a bipartition of the mechanism at the current state at time t, $\mathcal{M}_t$; $\mathcal{P}_{t\pm\tau}^1, \mathcal{P}_{t\pm\tau}^2$ refer to the blocks of a bipartition (not necessarily the same) of the present or past units $\mathcal{P}_{t\pm\tau}$. Figure 1B represents the partition $\mathcal{M}_1 = \{s_1(t), s_2(t)\}, \mathcal{M}_2 = \{s_3(t)\}, \mathcal{P}_1 = \{s_1(t+1), s_2(t+1), s_3(t+1)\}, \mathcal{P}_2 = \{\}$. The interaction between the partitioned systems ($\otimes$ operator) is defined by injecting uniform random noise in the partitioned connections when defining the transition probability matrix $P(\mathbf{s}(t \pm \tau)|\mathbf{s}(t))$.

The integrated information of the mechanism $\mathcal{M}_t$ with a time span τ, $\varphi(\tau)$, is the minimum of its corresponding integrated cause and effect information,

$$\varphi = \min(\varphi_{\text{cause}}, \varphi_{\text{effect}}). \tag{4}$$

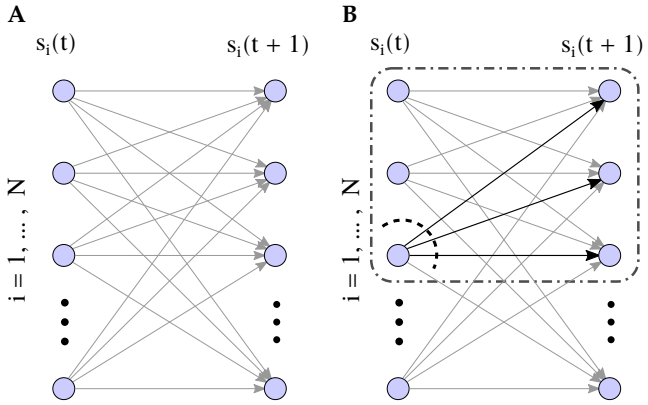

Figure 1. (**A**) Description of the infinite size kinetic Ising model. (**B**) Description of the partition schema used to define perturbations. Partitioned connections (black arrows) are injected with random noise. Nonpartitioned connections operate normally or are independent sources of noise (see Section 3.4).

The integrated information of the entire system, $\Phi(\tau)$, is then defined as the distance between the cause–effect structure of the system and the cause–effect structure of its minimum information partition, eliminating constraints from one part of the system to the rest:

$$\Phi(\tau) = \min_{\text{cut}} D_W^*(C(\tau), C_{\text{cut}}(\tau)), \tag{5}$$

where $C(\tau)$ stands for a "constellation of concepts", which is constructed from the set of points with position $\{P(\mathbf{s}(t-\tau)|\mathbf{s}(t)), P(\mathbf{s}(t+\tau)|\mathbf{s}(t))\}$ and value $\varphi(\tau)$ corresponding to all the mechanisms in the system. Similarly, $C_{\text{cut}}(\tau)$ stands for a constellation of a system in which a new unidirectional partition has been applied, injecting noise in the partitioned connections as in previous case (note that now φ is computed applying two different partitions). In this case, a especial distance measure is used (we label it as D_W^*), which is a modified he extended version of the Wasserstein distance that measures the minimal transport cost of transforming one constellation into another [1] (Text S2), in which $\varphi(\tau)$ are the values to be transported and the Wasserstein distance between $\{P(\mathbf{s}(t-\tau)|\mathbf{s}(t)), P(\mathbf{s}(t+\tau)|\mathbf{s}(t))\}$ in the original system and under the partition is the distance these values have to be transported. Finally, if the system is a subset of elements of a larger system, all elements outside the system are considered as part of the environment and are conditioned on their current state throughout the causal analysis. Similarly, when evaluating a mechanism, all elements inside a mechanism (where φ is analysed) but outside the system (where Φ is determined) are considered as uniform, independent noise sources. Further details of the steps described here can be found in [1].

Working Assumptions

In order to compute integrated information in large systems, we modify some aspects of the theory. In IIT 3.0, an integrated information of a mechanism φ is evaluated for a particular mechanism $\mathcal{M}_t$ and a purview $\mathcal{P}_{t\pm\tau}$. Here, for simplicity, we assume that the purview always includes the same units as the mechanism (although we allow them to be partitioned differently, see, e.g., Figure 1B). Allowing more options for the purview could make a big difference in some systems; although, in the homogeneous systems tested here, the differences are small. Also, the distance for computing integrated information is measured for the distance of all elements of the system, not only the elements contained in the purview. In IIT 3.0 only elements in the purview are affected by a partition. In our modified version of the measure, in some cases ($\tau > 1$, see Section 3.1) the outside of the purview can change as well, thus capturing these changes offers a better characterisation of the effects of partitions.

Moreover, as we assume a homogeneous architecture, in some cases mechanism integration φ and system-level integration Φ have a similarly diverging behaviour (as we explore in Section 3.5). Thus,

for simplicity, in most cases we compute only the integrated information φ of a mechanism comprising the system of interest.

This homogeneous architecture also allows us to assume that under some conditions (systems with possible couplings and near the thermodynamic limit) the MIP is always either a partition that cuts a single node from the mechanism of the system or a cut that separates entire regions in different partitions (see Appendix B.3). This assumption will reduce drastically the computational cost of calculating integrated information.

Other assumptions will be studied in different sections of the article. In Table A1, these different assumptions are described and in Table A2 it is indicated if they are used by IIT 3.0 and if they are applied for obtaining the results of the different figures of Section 3.

2.2. Kinetic Ising Model with Homogeneous Regions

We define a general model capturing causal interactions between variables. Aiming for generality, we use the least structured statistical model defining causal correlations between pairs of units from one time step to the next [15]. We study a kinetic Ising model where N binary spins $s_i \in \{-1, +1\}$ evolve in discrete time, with synchronous parallel dynamics (Figure 1A). Given the configuration of spins at time t, $\mathbf{s}(t) = \{s_1(t), \ldots, s_N(t)\}$, the spins $s_i(t)$ are independent random variables drawn from the distribution:

$$P(s_i(t)|\mathbf{s}(t-1)) = \frac{e^{\beta s_i(t) h_i(t)}}{2\cosh(\beta h_i(t))}, \tag{6}$$

$$h_i(t) = H_i + \sum_j J_{ij} s_j(t-1). \tag{7}$$

The parameters $\mathbf{H}$ and $\mathbf{J}$ represent the local fields at each spin and the couplings between pairs of spins, and β is the inverse temperature of the model. Without loss of generality, we assume that $\beta = 1$.

In general, computing the probability distributions $P(\mathbf{s}(t))$ of a kinetic Ising model is a difficult task, as it requires computing over all previous trajectories. In general, $P(\mathbf{s}(t+\tau)|\mathbf{s}(t))$ is computed recursively applying the equation

$$P(\mathbf{s}(t+\tau)|\mathbf{s}(t)) = \sum_{\mathbf{s}(t+\tau-1)} P(\mathbf{s}(t+\tau)|\mathbf{s}(t+\tau-1)) P(\mathbf{s}(t+\tau-1)|\mathbf{s}(t)). \tag{8}$$

The cost of calculating this equation grows exponentially with size of the system, with a computational cost of $\mathcal{O}(2^{2N})$.

This computation can be simplified for certain architectures of the model. We divide the system into different regions, and assume that the coupling values J_{ij} are positive and homogeneous for each intra- or inter-region connections $J_{ij} = \frac{1}{N_\mathcal{V}} J_{\mathcal{U}\mathcal{V}}$, where $\mathcal{U}$ and $\mathcal{V}$ are regions of the system with sizes $N_\mathcal{U}, N_\mathcal{V}$ and $i \in \mathcal{U}, j \in \mathcal{V}$. Also, for simplicity we assume that $\mathbf{H} = \mathbf{0}$.

When the system is divided in homogeneous regions, the calculation of the probability distribution of the system is simplified to computing the number of active units for each region $S_\mathcal{U}(t) = \sum_{i \in \mathcal{U}} (1 + s_i(t))/2$. With this, we simplify the transition probability matrix to

$$P(s_i(t)|\mathbf{S}(t-1)) = \frac{e^{\beta s_i(t) h_i(t)}}{2\cosh(\beta h_i(t))}, \quad h_i(t) = H_i + \sum_\mathcal{V} J_{\mathcal{U}\mathcal{V}} S_\mathcal{V}(t-1), \tag{9}$$

$$P(S_\mathcal{U}(t)|\mathbf{S}(t-1)) = P(s_i(t) = 1|\mathbf{S}(t-1))^{S_\mathcal{U}(t)} P(s_i(t) = 0|\mathbf{S}(t-1))^{(N_\mathcal{U} - S_\mathcal{U}(t))} \binom{N_\mathcal{U}}{S_\mathcal{U}(t)}. \tag{10}$$

Having then that

$$P(\mathbf{S}(t+\tau)|\mathbf{S}(t)) = \sum_{\mathbf{S}(t+\tau-1)} \prod_\mathcal{U} P(S_\mathcal{U}(t+\tau)|\mathbf{S}(t+\tau-1)) P(\mathbf{S}(t+\tau-1)|\mathbf{S}(t)). \tag{11}$$

Now the cost is reduced to $\mathcal{O}(\prod_{\mathcal{U}}(N_{\mathcal{U}}+1)^2)$, which, for a limited number of regions, makes the computation much lighter.

Interestingly, as shown in [14], if the size of the regions tends to infinity, then the behaviour of a region $\mathcal{U}$ is described simply by the mean field activity of its units, $m_{\mathcal{U}}(t) = \frac{1}{N_{\mathcal{U}}}\sum_{j\in\mathcal{U}} s_j(t)$, and its evolution becomes deterministic, having that the behaviour of any unit i belonging to a region $\mathcal{V}$ by the value of the input field of the region $h_{\mathcal{V}}$:

$$P(S_{\mathcal{U}}(t)|\mathbf{S}(t-1)) = \delta(S_{\mathcal{U}}(t) - N_{\mathcal{U}}\frac{1+m_{\mathcal{U}}(t)}{2}), \tag{12}$$

$$m_{\mathcal{U}}(t) = \tanh\sum_{\mathcal{V}} J_{\mathcal{U}\mathcal{V}} m_{\mathcal{V}}(t-1), \tag{13}$$

where $\delta(x)$ is the Kronecker delta function and $m_{\mathcal{U}}(t-1)$ is the mean field of region $\mathcal{U}(t-1)$.

2.3. Integrated Information in the Kinetic Ising Model with Homogeneous Regions

Describing an Ising model with homogeneous regions simplifies the computation of integrated information in two important ways: by reducing the costs of finding the minimum information partition and computing statistical distances between distributions.

As the connectivity of the system is homogeneous for all nodes in the same region, in Appendix B.3 we show that, near the thermodynamic limit and for the case of only positive couplings, the MIP is always a partition that either (a) isolates only one unit from one of the regions of the system, or (b) separates entire regions such that all elements of a region in the current or the future (or past) states always belong to the same partition. Also, in case (a), the partition that isolates a single unit in time t always has a smallest value of φ than the partition isolating a node at time $t \pm \tau$, as partitioning the posterior distribution corresponds to a larger distance between probability distributions (see Appendix B.3). We tested both cases (a) and (b) and found that in all the examples exemplified in this article, as all couplings are in a similar order of magnitude, the MIP always is as in case (a), so the MIP always cuts only a node at time t from one region of the system (see, e.g., Figure 1B). In this work, we also compute the value of Φ for a homogeneous system with just one region. In this case, as there is only one region, the MIP at the system level is also a partition that isolates only one node, as this is the intervention yielding a minimal distance (see Appendix B.4).

In systems with finite size, the evolution of the probability distribution of the activity at the regions of the system is calculated using Equation (11). From there, Equations (1) and (2) can be computed. For large systems, it becomes unfeasible to compute the Wasserstein distance between distributions due to the combinatorial explosion of states. Nevertheless, when regions are homogeneous this computation is greatly simplified if the number of regions is not too large. As all the units within a region are statistically identical, in terms of Wasserstein distances it is equivalent to work with aggregate variables representing the sum of the units of a region (see Appendix B.1):

$$\begin{aligned}\varphi^{\text{cut}}_{\mathcal{M}}(\tau) &= D_W(P(\mathbf{s}(\tau_0+\tau)|\mathbf{s}(\tau_0))||P_{\text{cut}}(\mathbf{s}(\tau_0+\tau)|\mathbf{s}(\tau_0)))\\ &= D_W((P(\mathbf{S}(\tau_0+\tau)|\mathbf{s}(\tau_0))||P_{\text{cut}}(\mathbf{S}(\tau_0+\tau)|\mathbf{s}(\tau_0))).\end{aligned} \tag{14}$$

This equivalence is possible because the transport cost between two states $\mathbf{s}(t)$ and $\mathbf{s}^*(t)$ is defined as $\frac{1}{2}\sum_i |s_i(t) - s_i^*(t)|$. As the Wasserstein distance always chooses minimal transport costs, then the cost between states $\mathbf{S}(t)$, $\mathbf{S}^*(t)$ is defined as $|\mathbf{S}(t) - \mathbf{S}^*(t)|$. If instead of the Wasserstein distance we use the Kullback–Leibler divergence, then $D_{KL}(\mathbf{s}(t)||\mathbf{s}^*(t)) = D_{KL}(\mathbf{S}(t)||\mathbf{S}^*(t))$ (see Appendix C).

Finally, when the partition only affects one node of the mechanism of the system at region $\mathcal{V}$ (see, e.g., Figure 1B), the computation of P_{cut} is performed by transforming the transfer probability matrix as

$$\begin{aligned} P_{\text{cut}}(S_{\mathcal{U}}(t)|\mathbf{S}(t-1)) = &\frac{1}{2}\Big(P(S_{\mathcal{U}}(t)|\mathbf{S}(t-1)) \\ &+ (1-\frac{S_{\mathcal{V}}(t)}{N_{\mathcal{V}}})P(S_{\mathcal{U}}(t)|S_{\mathcal{V}}(t-1)+1,\mathbf{S}_{\overline{\mathcal{V}}}(t-1)) \\ &+ \frac{S_{\mathcal{V}}(t)}{N_{\mathcal{V}}}P(S_{\mathcal{U}}(t)|S_{\mathcal{V}}(t-1)-1,\mathbf{S}_{\overline{\mathcal{V}}}(t-1))\Big), \end{aligned} \tag{15}$$

where $\overline{\mathcal{V}}$ is the complement set to $\{\mathcal{V}\}$. The origin of this expression is that injecting uniform noise to a single unit of region $\mathcal{V}$ can have three possible outputs: leaving the system as it was ($\frac{1}{2}$ chance), adding one to the value of $S_{\mathcal{V}}$ ($\frac{1}{2}(1-\frac{S_{\mathcal{V}}(t)}{N_{\mathcal{V}}})$ chance) or subtracting one to the value of $S_{\mathcal{V}}$ ($\frac{1}{2}\frac{S_{\mathcal{V}}(t)}{N_{\mathcal{V}}}$ chance). The linear combination of these three cases yields the final transfer probability matrix.

When computing integrated information of the system, Φ, Equation (5) computes the distance between concepts (i.e., values of integration of the mechanisms of a system) of the original system and the system under a unidirectional partition. Mechanisms affected by the partition will always have a value of $\varphi = 0$ (and are transported to a residual, "null" concept located at the unconstrained distribution, see [1] (Text S2)). In our example, we find that these concepts contribute the most to the value of Φ (see Section 3.5).

3. Results

Criticisms concerning the definition of integrated information measures have addressed a variety of topics, e.g., the existence of "trivially non-conscious" systems, composed of units distributed in relatively simple arrangements yielding arbitrarily large values of Φ, or the absence of a canonical metric of the probability distribution space [5]. There are also some aspects that are not very well understood as the dependence of Φ with the scale or graining of a system, or the differences and dependencies between cause and effect repertoires. Here, we explore some of these aspects and explore possible reformulations of the measures concerning how current measures behave around critical phase transitions. We introduce as a reference the behaviour of the kinetic Ising model with homogeneous regions of infinite size. As described in Equation (13), behaviour in the thermodynamic limit can be described by the evolution of the mean firing rates. Also, computing the derivative of the mean firing rates with infinite size (which will be used to compute the distances between distributions) is straightforward:

$$\frac{\partial m_{\mathcal{U}}(t)}{\partial J_{\mathcal{U}\mathcal{V}}} = (1-m_{\mathcal{U}}^2(t))(m_{\mathcal{U}}(t-1)+\sum_{\mathcal{V}} J_{\mathcal{U}\mathcal{V}}\frac{\partial m_{\mathcal{V}}(t-1)}{\partial J_{\mathcal{U}\mathcal{V}}}). \tag{16}$$

In Figure 2 we observe an example for a infinite-size kinetic Ising model with just one homogeneous region $\mathcal{U}$ with self couplings of value $J_{\mathcal{U}\mathcal{U}} = J/N$. In this case, the model displays a critical point at $J = 1$.

We argue that this critical phase transitions offers an interesting case for studying integrated information. First, systems at criticality display long-range correlations and maximal energy fluctuations and at the critical point [7,8], which should produce maximal dynamical integration, as noted in [16]. However, IIT 3.0 is concerned not with dynamics, but with causal interactions (i.e., how the states of mechanisms generate information by constrain future/past states), thus assuming critical dynamics is not enough for expecting maximum integration in the terms of IIT 3.0. Still, we can argue that (a) phase transitions in an Ising model mark a discontinuity between different modes of operation and (b) the critical point is characterised by maximum susceptibility (i.e., sensitivity to changes in intensive physical properties of mechanisms, e.g., Figure 2B) in front of external perturbation [7]. Because of this, when measuring integrated information in Ising models, we expect

critical phase transitions to be observable in terms of integrated information, and critical points to have distinguishable properties respect other points of the phase space.

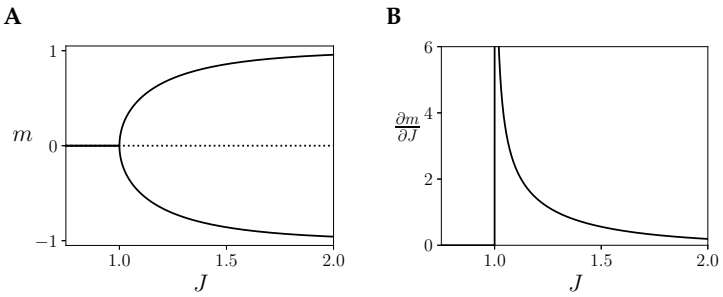

Figure 2. Description of the behaviour of the homogeneous Ising model with one region and coupling J, showing a critical point at $J = 1$. (**A**) Values of mean firing rate m for the stationary solution of the kinetic Ising model with one homogeneous region. (**B**) Value of $\frac{\partial m}{\partial J}$ for the positive stationary solution of the kinetic Ising model with one homogeneous region, diverging at the critical point.

Using this toy model, we explore different aspects of the mathematical definitions of integrated information and the assumptions behind these definitions, using critical phase transitions as a reference. We will compute $\varphi(\tau)$ as follows. First, we select the initial state $\mathbf{s}(t)$. For finding a representative initial state, we start from a uniform distribution of $P(\mathbf{s})$ (and the corresponding $P_{\mathrm{I}}(\mathbf{S}) = \frac{1}{2^N}\binom{N}{S}$) and update until it stabilizes (using Equation (11) or Equation (13) for the finite and infinite cases, respectively). Then, we choose $\mathbf{S}(t) = \arg\max P(\mathbf{S})$. From there, we update the probability distributions forward or backwards τ times with and without applying a partition and compute the distance between distributions for computing φ_{effect} and φ_{cause}. Total integrated information φ will be computed as the minimum between the two. In all sections, we will assume that mechanism and purview contain the same units $\{i\}_{i\in I_{\mathcal{M}_t}} = \{i\}_{i\in I_{\mathcal{P}_{t\pm\tau}}}$. We also will assume that the mechanism and the purview are composed by the whole system under analysis, except for Sections 3.4 and 3.5, were smaller subsystems are analysed. Only in Section 3.5 do we compute the value of Φ, and we will argue that in our examples the first level of integration φ is enough for describing the behaviour of the system.

3.1. Dynamics and Temporal Span of Integrated Information

First, we explore integrated information of the effect repertoire of a system for a time span τ, $\varphi_{\text{effect}}(\tau)$. In IIT 3.0, integrated information is defined as the level of irreducibility of the causal structure of a dynamical system over one time step [1]. The level of irreducibility is computed by applying partitions over the system, in which noise is injected in the connections affected by the partition. This is done by manipulating the transition probability matrix of the system. In previous work, integrated information has been applied to different temporal spans by changing the temporal graining of the system and joining consecutive states in a new Markov chain [12] or by concealing the micro levels in black box mappings [17]. Another possibility could be describing the transition probability matrix of the system from t and $t + \tau$. However, is this the adequate way to capture integration at larger timescales? As IIT 3.0 operates with the transition probability matrix of a system, one could compute this matrix from time t to time $t + \tau$ and compute a new transition probability matrix for a bipartition by injecting noise in the connections affected by it at time t. This implies that noise is injected at the first step (at time t) and then the system behaves normally for the following steps. We will refer to this way of applying a partition as an 'initial noise injection' (in contrast with a "continuous noise injection", see below).

We explore this by computing integrated information with only one region of size $N = 256$, with coupling values $J_{ij} = J/N$. If we compute φ for different values of τ (Figure 3A), we observe that for different couplings J integrated information always peaks at the ordered side of the phase transition.

As τ is increased, this peak moves towards the critical point and its size decreases decreases, tending to zero. The assumption of an initial noise injection yields $\varphi(\tau) = 0$ at the critical point and maximum integration at the ordered side of the phase transition. Thus, integrated information in this case is not able to characterise the phase transition of the system

A different metric can be defined if, instead of applying the partition just at the initial step, we apply it to all τ updates (Figure 3B). We will refer to this way of applying a partition as a "continuous noise injection", in contrast with the case in which noise is only injected at the first step. We propose that this is a more natural way to apply a partition, capturing larger integrated information around the critical point as we consider larger timescales. Moreover, as opposed to the previous case, which captured zero integration in the disordered side, this measure is able to capture increasing integration as the system approaches the critical transition from any side. One may note that in the mean-field approximation for infinite size (as shown in [14] and also Figure 5A), integration is zero when approaching a critical point from the disorder side. This is not a problem of the measure but a characteristic of the system, in which units have independent dynamics until J reaches the threshold of the critical point. For finite size, units are not completely independent and our measure correctly captures non-zero integration.

Still, some important considerations need to be taken into account when applying a continuous noise injection. In a initial noise injection, φ decreases with time as the effect of causal structures is diluted with time. In contrast, a continuous noise injection accumulates the effects of each time step, making integration grow for larger temporal span. These are very different assumptions, but we propose that the latter is more appropriate in our case in order to capture the long-range correlations and critical slowing down properties displayed by systems at criticality. Therefore, for the remainder of the article, we will assume a continuous noise injection.

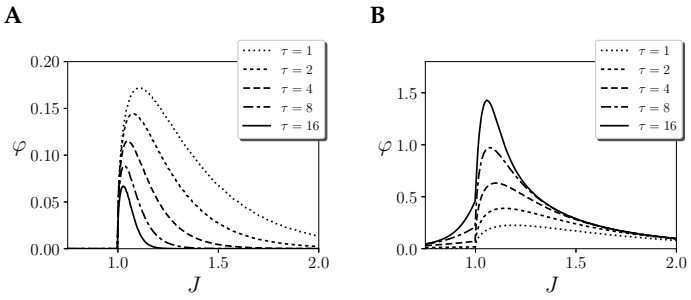

Figure 3. Integration of the effect repertoire $\varphi_{\text{effect}}(\tau)$ of the largest mechanism of a homogeneous Ising model with one region of size $N = 256$ and couplings J with different temporal spans τ, assuming (**A**) initial injection of noise and (**B**) continuous injection of noise. Note that $\tau = 1$, in both cases φ_{effect}, has the same value.

3.2. Integrated Information of the Cause Repertoire

In the previous section we have explored the behaviour of φ_{effect} around a critical phase transition, i.e., the value of integrated information for the repertoire of states generated by the mechanisms of a system at time $t + \tau$. IIT 3.0 proposes that integrated information should be computed as the minimum between φ_{effect} and φ_{cause} (Equation (4)). This is motivated by the "intrinsic information bottleneck principle", proposing that information about causes of a state only exist to the extent it also can specify information about its effects, and vice versa [1].

Describing the cause repertoire is more complicated than the effect repertoire. IIT 3.0 [1] (Text S2) proposes to tackle the problem of defining $P(\mathbf{s}(t-1)|\mathbf{s}(t))$ by assuming a uniform prior distribution of past states $P_U(\mathbf{s}(t-1))$. This takes the form

$$P(\mathbf{s}(t-1)|\mathbf{s}(t)) = \frac{P(\mathbf{s}(t)|\mathbf{s}(t-1))P_{\mathrm{U}}(\mathbf{s}(t-1))}{\sum_{\mathbf{s}(t-1)} P(\mathbf{s}(t)|\mathbf{s}(t-1))P_{\mathrm{U}}(\mathbf{s}(t-1))}, \tag{17}$$

where P_U stands for a uniform probability distribution. This is equivalent to

$$P(\mathbf{S}(t-1)|\mathbf{S}(t)) = \frac{P(\mathbf{S}(t)|\mathbf{S}(t-1))P_{\mathrm{I}}(\mathbf{S}(t-1))}{\sum_{\mathbf{S}(t-1)} P(\mathbf{S}(t)|\mathbf{S}(t-1))P_{\mathrm{I}}(\mathbf{S}(t-1))}, \tag{18}$$

where $P_{\mathrm{I}}(\mathbf{S}) = \frac{1}{2^N}\binom{N}{S}$ is the binomial distribution resulting from combining N independent distributions (obtained directly from $P_U(\mathbf{s})$). Similarly, $P_{\mathrm{cut}}(\mathbf{s}(t-1)|\mathbf{s}(t))$ and $P_{\mathrm{cut}}(\mathbf{S}(t-1)|\mathbf{S}(t))$ can be computed like in Equations (17) and (18), assuming a modified conditional probability $P_{\mathrm{cut}}(\mathbf{S}(t)|\mathbf{S}(t-1))$

What is the effect of considering an independent prior? As we observe in Figure 4A, for $\tau = 1$, integration remains high even for large values of J. As we increase τ, integration decreases. This behaviour is completely different to the effect repertoire (Figure 3B). Intuitively, such a difference of behaviour from cause and effect mechanisms is strange for an homogeneous system in a stationary state as the one under study here. More importantly, the measure of φ_{cause} fails to capture integration around the critical point, and displays the largest values of integration of the system far into the ordered side of the phase space. Note that as $\varphi = \min(\varphi_{\mathrm{cause}}, \varphi_{\mathrm{effect}})$, in this case, the value of integration would be dominated by the cause repertoire, and φ would not diverge around the critical point.

It is possible to drop the assumption of an independent prior, but some assumption about the prior distribution is needed. A simple alternative is to assume that the system is in a stationary state with distribution $P_{\mathrm{st}}(\mathbf{s}(t)) = P_{\mathrm{st}}(\mathbf{s}(t-1))$, having then

$$P(\mathbf{s}(t-1)|\mathbf{s}(t)) = \frac{P(\mathbf{s}(t)|\mathbf{s}(t-1))P_{\mathrm{st}}(\mathbf{s}(t-1))}{\sum_{\mathbf{s}(t-1)} P(\mathbf{s}(t)|\mathbf{s}(t-1))P_{\mathrm{st}}(\mathbf{s}(t-1))} = \frac{P(\mathbf{s}(t)|\mathbf{s}(t-1))P_{\mathrm{st}}(\mathbf{s}(t-1))}{P_{\mathrm{st}}(\mathbf{s}(t))}, \tag{19}$$

$$P(\mathbf{S}(t-1)|\mathbf{S}(t)) = \frac{P(\mathbf{S}(t)|\mathbf{S}(t-1))P_{\mathrm{st}}(\mathbf{S}(t-1))}{\sum_{\mathbf{S}(t-1)} P(\mathbf{S}(t)|\mathbf{S}(t-1))P_{\mathrm{st}}(\mathbf{S}(t-1))} = \frac{P(\mathbf{S}(t)|\mathbf{S}(t-1))P_{\mathrm{st}}(\mathbf{S}(t-1))}{P_{\mathrm{st}}(\mathbf{S}(t))}, \tag{20}$$

In this case, computing $\varphi_{\mathrm{cause}}(\tau)$ (Figure 4B), we observe that the integration of the cause and effect repertoires has a similar behaviour as J changes, yielding similar curves to $\varphi_{\mathrm{effect}}(\tau)$. Still, note that integration values are slightly lower for the cause repertoire.

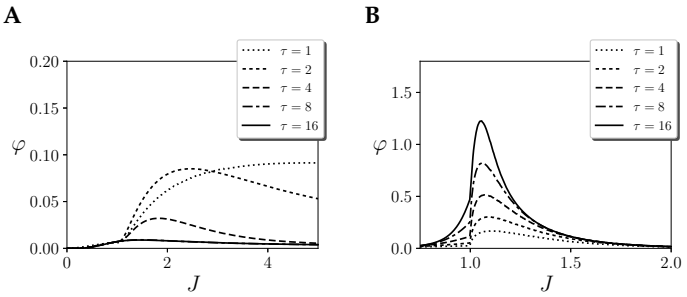

Figure 4. Integration of the cause repertoire $\varphi_{\mathrm{cause}}(\tau)$ of the largest mechanism of a homogeneous Ising model with one region of size $N = 256$ and couplings J with different temporal spans τ, assuming (**A**) an independent prior and (**B**) the stationary distribution as a prior. Continuous noise injection is assumed.

Thus, the assumption of an independent prior has dramatic consequences, which can be avoided by assuming an stationary distribution. Another alternative for systems undergoing a transient is to compute the trajectory of probability distributions $P(\mathbf{s}(t))$ and use it as priors, though this makes the computation much more costly. For the rest of the manuscript, we will assume an stationary prior. Note that the noise injected when partitioning the system is still uniform in all case.

3.3. Divergence of Integrated Information: Wasserstein and Kullback–Leibler Distance Measures

We have seen that, using our assumptions, φ grows with τ around the critical point in a finite system, suggesting that the value of integration may diverge in the thermodynamic limit. We test this divergence by computing integrated information φ for for networks of different size N and a given τ. In general, the relationship of φ and τ is complex, as for each value of J, it depends on the transient dynamics of the system. It is not the goal of this article to explore this issue in detail, but we want to ensure that finite systems have enough time to get close to a stationary regime. Thus, from now on, for simplicity, we will use a value of $\tau = 10 \log_2 N$, where N is the size of the system. We choose this relation because we have tested that it ensures the divergence of integrated information around critical points, although other relations we tested (e.g., $\tau \propto N$) maintain the qualitative results shown in the following sections.

To test the divergence of φ, we compute the value of integrated information for the largest mechanism of a kinetic Ising model with an homogeneous region with different size N, and assuming continuous noise injection and a stationary prior. We observe in Figure 5A that for finite sizes φ_{cause} (black line) shows a diverging tendency around the critical point. Effect integration φ_{effect} (grey line) shows a similar divergence, with values slightly larger. In this case, we also computed the value of φ_{effect} for infinite size. When $N \to \infty$, units $s_i(t+\tau)$ become independent, and the Wasserstein distance of a system with one region is computed analytically as $D_W(P(\mathbf{s}(t+\tau)|\mathbf{s}(t+\tau))||P_{\text{cut}}(\mathbf{s}(t+\tau)|\mathbf{s}(t+\tau))) = \frac{1}{2}\frac{\partial m(t+\tau)}{\partial J} J$ (see Appendix B.2). The divergence of φ_{effect} for infinite size shown in Figure 5A was also analytically characterised in [14]. As φ_{effect} is always larger than φ_{cause}, in this case the total integration is always $\varphi = \varphi_{\text{cause}}$. Summarising, we can conclude that φ computed using the Wassertein distance shows a divergence around the critical point of the kinetic Ising model.

Many versions of φ use the Kullback–Leibler divergence as an alternative to the Wasserstein distance. As seen in Figure 5B, this change can lead to an important difference in the results of φ_{cause} (black line) and φ_{effect} (grey line). The figure shows that φ tends to peak around the critical point but decreases with the size of the system. Also for this case, $\varphi = \varphi_{\text{cause}}$ for the cases we computed. By doing a similar approximation than in the previous case can be used to compute φ_{effect} in the infinite case (see Appendix C), using the well-known relation between the Kullback–Leibler and Fisher information, it can be shown that $D_{KL}(P(\mathbf{s}(t+\tau)|\mathbf{s}(t+\tau))||P_{\text{cut}}(\mathbf{s}(t+\tau)|\mathbf{s}(t+\tau))) = \frac{1}{2}\frac{1}{1-m^2(t+\tau)} J^2 \frac{1}{N}\left(\frac{\partial m(t+\tau)}{\partial J}\right)^2$, tending to zero for diverging size N (see Appendix C). Using this expression, we find that for infinite size the value of $\varphi_{\text{effect}} N$ diverges. However, computing the values for the finite networks, we find that $\varphi_{\text{effect}} N$ and $\varphi_{\text{cause}} N$ do not diverge for finite values of N (Figure 5C). This can be interpreted as a similar phenomenon found in homogeneous Ising models (e.g., Curie–Weiss model [13]), where the Heat Capacity, equivalent to the Fisher approximation to the Kullback–Leibler divergence computed here, does not diverge for finite sizes as the size of the system grows.

These results illustrate that different distance measures can have important effects in the behaviour of φ. As well, our results show that different metrics can hold different relations between finite models and the mean-field behaviour of the model with infinite size. For the Wasserstein distance, φ in finite systems tends to a diverging behaviour around the critical point, characterised for the infinite mean-field model. Conversely, for the Kullback–Leibler divergence, φ does not diverge in finite models and it does diverge for the mean-field infinite model (for the effect repertoire). In this case, the symmetry breaking of the system for infinite size provokes that the behaviour of the system is different in the mean-field model (a similar phenomena takes place with simple measures as the

average magnetisation). This effect can be relevant for studying φ in real finite systems by computing their mean-field approximations.

Although most versions of integrated information measures used the Kullback–Leibler divergence, recently IIT 3.0 suggested that the Wasserstein distance is a more appropriate measure [1] (Text S2). The result presented here show that the change of distance measure can have important implications when measuring large systems. Further work should inspect how different distance measures are able or not to capture the scaling behaviour of different systems and how is this coherent with the properties of the systems under study. A way to do so could be to explore the connection between the relation of the Wasserstein and Kullback–Leibler versions of integrated information with well-known variables in Ising models like the magnetic susceptibility or the heat capacity of the system (see Appendices B and C).

As we have shown that under the appropriate assumptions both φ_{cause} and φ_{effect} have similar diverging tendencies, in the rest of the example, we will not show these variables separately and will just show $\varphi = \min(\varphi_{\text{cause}}, \varphi_{\text{effect}})$.

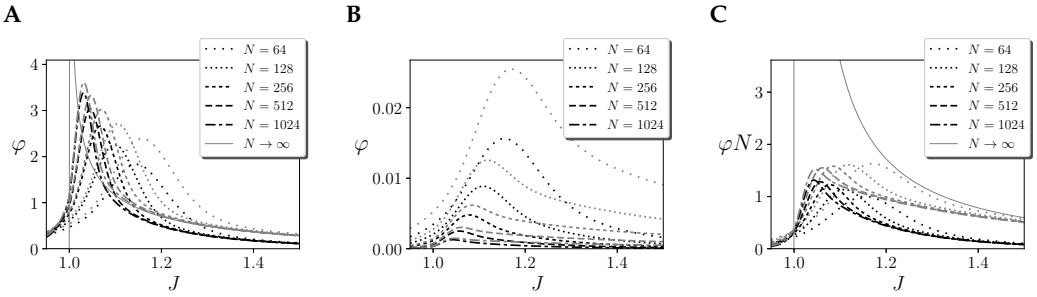

Figure 5. Integrated information $\varphi(\tau)$ for the cause (black lines) and effect (grey lines) repertoires of the largest mechanism of a homogeneous kinetic Ising models with one region of size N (and infinite size when $N \to \infty$) and coupling J using (**A**) the Wasserstein distance. (**B**) The Kullback–Leibler divergence, and (**C**) values of φN using the Kullback–Leibler divergence. Note that in all cases $\varphi(\tau) = \varphi_{\text{cause}}(\tau)$. All cases are computed with $\tau = 10 \log_2 N$ for finite systems and $\tau \to \infty$ for infinite systems. Continuous noise injection and a stationary prior are assumed.

3.4. Situatedness: Effect of the Environment of a System

In IIT 3.0, there is a difference between a *mechanism*, where a first level of integration φ is computed, and a *candidate set*, composed of different mechanisms, where a second-order level of integration Φ applies. When computing integration at these two levels, there are some assumptions about how the elements outside of the systems are considered. In IIT 3.0, the elements inside the candidate set but outside of the mechanism are treated as independent sources of noise. Respectively, elements outside the candidate set are treated as background conditions, and are considered as fixed external constraints.

What are the effects of these assumptions when computing integrated information in a critical phase transition? We measure again integration of a kinetic Ising model with one region of size N and coupling J. However, instead of considering the whole system, we measure the level of integration of a subsystem or mechanism $\mathcal{M}$ covering a fraction of the system M/N, where M is the size of the mechanism. We choose a value of $M = \frac{3N}{4}$, although other fractions yield similar results. To compute Equation (11) with and without the partition, we divide the system in two regions: one consisting on the units belonging to the mechanisms, and the other containing the units outside the mechanism. We measure the integrated information of the mechanism $\varphi_{\mathcal{M}}$ under three different assumptions: (a) that units outside of the mechanism operate normally, (b) units outside the mechanism are independent noise sources and (c) units outside the mechanism are fixed as external constraints.

In the first case, when external units operate normally (Figure 6A), we observe that the divergence of $\varphi_{\mathcal{M}}$ is maintained (although testing different values of M shows that $\varphi_{\mathcal{M}}$ increases with the

size of the mechanism, see [14]). In contrast, if we accept the assumptions of IIT 3.0 and take the elements outside the mechanism as independent sources of noise or as static variables, the behaviour of $\varphi_{\mathcal{M}}$ changes radically. In the former case, when outside elements are independent noise sources the divergence is maintained but takes place at a different value of the parameter J (Figure 6B). This happens because inputs from uniform independent sources of noise will be distributed around a zero mean field value, and thus the phase transition of the system takes place at larger values of J that compensate for the units that are now uncorrelated. Thus, considering the elements outside of the mechanism as independent sources of noise can be misleading, showing that maximum integration takes places at different points of the system. In this case, the position of the divergence is located at larger values of J, corresponding with significantly lower values of covariance and fluctuations in the units of the system, therefore not reflecting the actual operation of the mechanisms.

The latter assumption implies maintaining the state of the units outside of the mechanism with the static values that they had at time t. In this case (Figure 6C), we find that $\varphi_{\mathcal{M}}$ does not diverge, and instead it has a peak in the ordered side of the phase transition. We can understand this by thinking that the effect of constant fields is equal to adding a value of H_i equal to the input from static units, therefore breaking the symmetry of the system and precluding a critical phase transition. In both cases, we observe that ignoring or simplifying the coupling between a system and its environment can affect drastically the values of integration as a system scales.

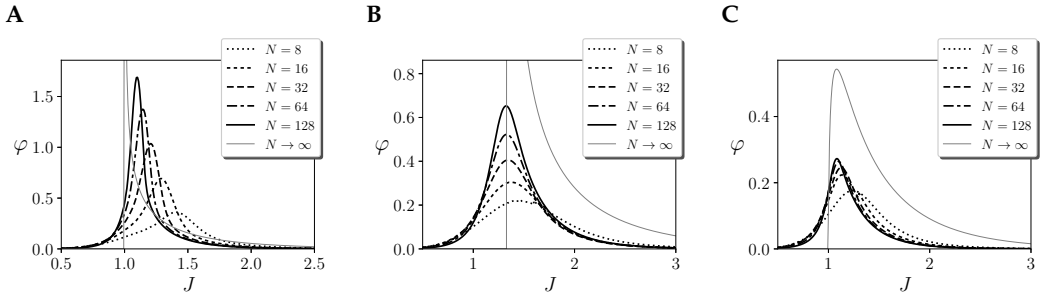

Figure 6. Effects of the environment in integrated information. Integrated information $\varphi_{\mathcal{M}}(\tau)$ (black lines) of a mechanism $\mathcal{M}$ of size $\frac{3N}{4}$ of a homogeneous kinetic Ising model with one region of size N and coupling J, assuming that elements outside of the mechanism operate (**A**) normally, (**B**) as independent sources of noise and (**C**) as static input fields. Values of $\varphi_{\mathcal{M}}(\tau)$ are compared with $\varphi_{\mathcal{M},\text{effect}}(\tau \to \infty)$ (grey line) to show diverging tendencies of the effect repertoire. Note that tendencies of $\varphi_{\mathcal{M},\text{effect}}(\tau \to \infty)$ are larger than values of $\varphi_{\mathcal{M}}(\tau)$, as the effect repertoire tends to show larger values. Values of φ are computed with $\tau = 10 \log_2 N$ for finite systems and $\tau \to \infty$ for infinite systems. Continuous noise injection and a stationary prior are assumed.

3.5. System-Level or Mechanism-Level Integration: Big Phi versus Small Phi

So far, we analysed the behaviour of integration measures φ describing the integration of mechanisms of a system. IIT 3.0 postulates that the integration of a system is defined by a second-order measure of integration, Φ, applied over the set of all its mechanisms. To explore how Φ behaves for systems of different sizes, we compute it for a homogeneous system with one region, including the different modifications assumed in the previous subsections. Note that this modifies the measure, but it still allows us to inspect some of its scaling aspects.

For measuring Φ, first φ is computed for the different mechanisms of the system, and then the integration of the set of mechanisms is compared with the set of values of φ of the system under unidirectional partitions, using a modified Wasserstein distance [1] (Text S2). In the case of a homogeneous system with just one region, the MIP is any of the partitions that isolates one single node from the rest of the system. The value of Φ is the modified Wasserstein distance (with and without applying the MIP) between the values of φ of the set of mechanisms of the system.

In Figure 7, we compare the values of φ of the larger mechanism (Figure 7A) with the normalised values of Φ of the whole system (Figure 7B), for a homogeneous system with only one region with self-couplings J. The value of φ of the larger mechanisms diverges around the critical point as expected. In the case of Φ, we find that for all values of J integration grows with size very rapidly. This is due to the fact that the number of concepts (i.e the number of mechanisms) of the system grows exponentially with size. The number of mechanisms or concepts is $N_C = \sum_{k=1}^{N} \binom{N}{k} = 2^N - 1$. Thus, we normalise the value of Φ dividing by N_C (Figure 7B). Using normalised values of Φ, we observe that the system still diverges at the critical point. Furthermore, in this case the divergence is faster than in the case of φ, as it accumulates the effects of the divergence of many mechanisms under a second partition.

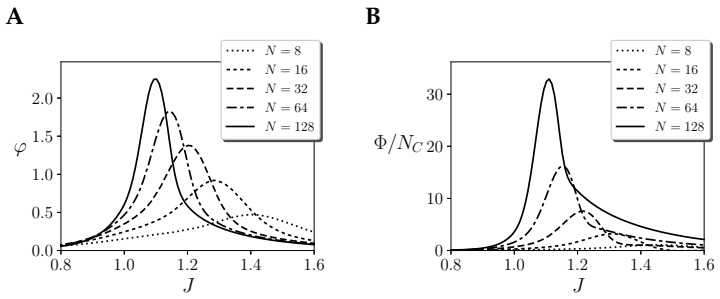

Figure 7. Mechanims and system-level integration in a homogeneous system with one region of size N and coupling J. Values of (**A**) φ of the largest mechanism and (**B**) values of Φ for the whole system. Measures with $\tau = 10 \log_2 N$, assuming continuous noise injection, stationary priors and environment coupling.

If we observe the contribution of different mechanisms to Φ, we observe that most of the contributions to Φ are determined by the mechanisms affected by the MIP. In this case, all the value of φ is transported by the Wasserstein distance into a new point defined by an independent distribution [1] (See Text S2) (e.g., for $N = 128$ around 98% of the value of Φ is defined by the value of φ of the mechanisms under the MIP).

In our example, it seems that the relation between φ and Φ is not quite relevant (the divergence of the later seems to be an amplified version of the former). Heterogeneous or sparsely connected systems may present more complicated relations and present important differences in the behaviour of the highest order φ and the total Φ. Still, we believe that our simple example calls for a better justification of the need of measuring a second-order level of integration in IIT 3.0 and the difference of the two levels respect well-studied properties of systems.

3.6. Values versus Tendencies of Integration

Finally, we explore mechanism integration φ in the case of two homogeneous regions: one region $\mathcal{A}$ with self-interaction and another region $\mathcal{E}$, which is just coupled to the first without recurrent connections (i.e., $J_{\mathcal{E}\mathcal{E}} = 0$, Figure 8A). This case was used in [14] to represent an agent interacting with an environment, exploring the power of integrated information to delimit what is the most integrated part of the system. This delimitation has been proposed to identify the autonomy of small biological circuits [2], but it is still unclear if the conclusions of analysis in such small systems and simplified models could be extended to larger models of neural and biological systems.

For different values of recurrent connections $J_{\mathcal{A}\mathcal{A}} = J_R$, two values of bidirectional couplings $J_{\mathcal{A}\mathcal{E}} = J_C$, $J_{\mathcal{E}\mathcal{A}} = 2J_C$ are tested: $J_C = 0.8$ and $J_C = 1.2$. The results in [14] showed that, for infinite sizes, in the weaker coupling condition $J_C = 0.8$, $\mathcal{A}$ was the most integrated unit of the system at the critical point. In contrast, for a stronger coupling $J_C = 1.2$, the joint $\mathcal{A}\mathcal{E}$ system was the one that presented higher integration for infinite size. In Figure 8, we show the values of integration of $\mathcal{A}$ and $\mathcal{A}\mathcal{E}$ for

different sizes (integration of $\mathcal{E}$ is always zero as there are no recurrent connections for this region), with $J_C = 0.8$ (Figure 8B,C) and $J_C = 1.2$ (Figure 8D,E).

For $J_C = 0.8$, we observe that $\varphi_{\mathcal{A}}$ is always larger, independently of the size of the system, showing that $\mathcal{A}$ is always more integrated than $\mathcal{AE}$. However, for $J_C = 1.2$, we find an interesting behaviour. We can observe that for small sizes ($N = 8, 16$) $\mathcal{A}$ is more integrated. Conversely, for larger sizes ($N = 64, 128$) we observe that $\mathcal{AE}$ is more integrated, as its value of $\varphi_{\mathcal{A}}$ diverges faster with size than $\varphi_{\mathcal{AE}}$.

This is relevant because in many cases integrated information can only be measured for rather small systems. When analysing models of real neural or biological systems, these should be coarse grained or discretised in order for φ measures to be applicable. In such cases, we can expect that the delimitation of the most integrated units of the system have different values than at larger scales. Thus, rather than the exact value of φ, the diverging tendencies in the model might be most informative about the behaviour of the real observed system when small networks are considered.

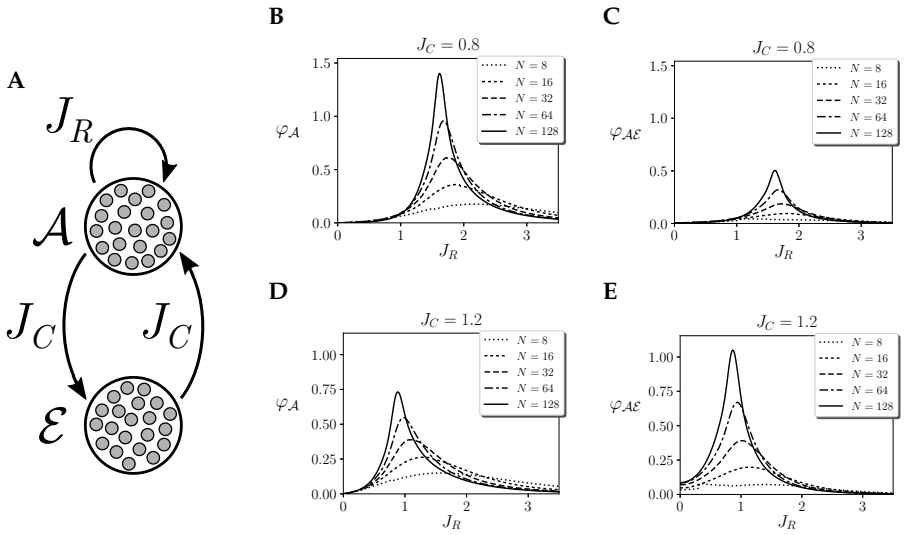

Figure 8. Integrated information in a system coupled to an environment. (**A**) Structure of couplings between the two regions $\mathcal{A}, \mathcal{B}$ of size $N_{\mathcal{A}} = N_{\mathcal{E}} = \frac{N}{2}$ of a homogeneous kinetic Ising models with couplings $J_{\mathcal{AA}} = J_R, J_{\mathcal{EE}} = 0, J_{\mathcal{AE}} = J_C, J_{\mathcal{EA}} = 2J_C$. (**B**–**E**) Integrated information of the mechanism $\mathcal{A}$, $\varphi_{\mathcal{A}}$ and mechanism $\mathcal{AE}$, $\varphi_{\mathcal{AE}}$, for values of $J_R = 1$ and $J_C = 0.8$ and $J_C = 1.2$, respectively. Values of φ are computed for $\tau = 10 \log_2 N$ for finite systems and $\tau \to \infty$ for infinite systems. Continuous noise injection, stationary priors and environment coupling are assumed.

4. Discussion

In this article, we critically reviewed different aspects of the definition of integrated information proposed by IIT 3.0, exemplifying them in toy models displaying critical phase transitions. Using a homogeneous Ising model, we simplify the calculations to measure integrated information in large systems. It is well known from theory in spin glasses that the infinite range homogeneous Ising model (also known as Curie Weiss model) presents a critical point for $\mathbf{J} = \mathbf{1}$ and $\mathbf{H} = \mathbf{0}$ [13]. Although we argue that critical phase transitions should be observed in integrated information measures (as critical points display long-range correlations, maximal susceptibility to parametric changes and preserve integrative and segregative tendencies of the system, see [7,16]), we have shown how different aspects of the definition of φ prevent to capture the critical phase transition of the system as it grows larger in size. This investigation has led us to propose reformulations of some aspects of the theory in order to address some of the problems encountered during the study.

As IIT 3.0 has been mostly tested in small logic gate circuits, exploring the behaviour of integrated information in large Ising models has allowed us to investigate questions that were so far unexplored

and inspect some of the assumptions of the theory from a new perspective. We consider that the value of the study is twofold. On one hand, we propose a family of models with known statistical properties, where calculations of integrated information are simplified. These and similar models could work as a benchmark for testing properties of integrated information in large systems. On the other hand, the reformulations of different aspects of the theory proposed during the paper could be considered by future versions of IIT, to capture some of the phenomena that we could expect in large, complex systems.

First, we explored how the application of integrated information over an adequate timescale is important to observe increasing values of integration as the system scales. The dynamics of the Ising model are characterised by a "critical slowing down" as the critical point is approached. Consequently, we observed that to capture critical diverging tendencies, timescales larger than one time step should be used. As the dynamics of critical system display correlations at very different timescales, and the span of these timescales increase with the size of the system, integrated information should be evaluated in a way that the diversity of timescales is captured. In our analysis, we found that the way to capture integration near critical points is to apply partitions in a different way than IIT 3.0. In IIT 3.0 partitions are applied by injecting noise in the input state of the system and then computing the forward and backwards distributions, but this approach did not capture the phase transition in the model. In contrast, we successfully characterised the phase transition as diverging integrated information around the critical point by applying several updates of the state of the system and injecting noise at each update.

Second, to capture the cause repertoire of a state (integration in the causal dependencies of a mechanism with previous states), IIT 3.0 proposes to assume a uniform prior distribution of past states. We show that this assumption can distort the observed values of integration, losing an adequate characterisation of the critical phase transition. We suggest that the real prior distribution (either stationary or transient) should be used if cause repertoires are considered.

The third aspect we studied is the use of different distance measures between probability distributions. Specifically, we compared the Wasserstein distance used by IIT 3.0 with the Kullback–Leibler divergence, which is the choice for many competing definitions of integrated information. First, we show that values of the Kullback–Leibler divergence should be weighted by the size of the system in order to be comparable to the Wasserstein distance under the MIP; otherwise, they tend to zero as the system grows. We also show that, in a homogeneous kinetic Ising model at criticality, the Wasserstein distance shows diverging tendencies for finite sizes, whereas the Kullback–Leibler divergence only shows a finite peak. This shows that, in some cases, the Wasserstein distance may detect some divergences that would be ignored by the Kullback–Leibler divergence. Still, it should be debated whether it is adequate that a system like the toy model presented here shows a diverging value of integration. A closer examination of the behaviour of known quantities in an Ising model could constitute an adequate starting point for this discussion. In this sense, the results of the Wasserstein distance and Kullback–Leibler divergence can be connected with the behaviour of known quantities in the homogeneous Ising model. For example, the susceptibility of the system diverges at the critical point while the heat capacity of the system only shows a peak [13]. Both measures can be related to approximations of φ using the Wasserstein and Kullback–Leibler measures, respectively (from Equations (A22) and (A9)).

Furthermore, we analysed a crucial aspect of integration measures that is often overlooked: the situatedness of the system. The central claim of situated approaches to cognitive behaviour is that the agent-environment coupling shapes brain dynamics in a manner that is essential to behavioural or cognitive functionality [18,19]. Thus, ignoring or dismissing this brain-body-environment coupling can result in a substantial quantitative and qualitative distortion of the activity of a neural system [20]. Besides, there are deep theoretical reasons that come from the enactive perspective on cognition that establish that the very notion of cognitive performance is world-involving, i.e., that it is co-constituted by agent and environment [21]. In contrast, IIT 3.0 dismisses the bidirectional interaction between

the system under evaluation and its environment for computing integration, with the aim to assess the integrated information from the "intrinsic perspective" of the system itself. Specifically, IIT 3.0 considers the units outside the system (i.e., outside the candidate set) as static variables and the units within the system but outside the evaluated mechanism as independent sources of noise. We show in the model that both assumption can have dramatic effects in the behaviour of the system. The assumption of static variables makes the divergence at the critical point disappear, and the assumption of independent sources of noise creates spurious divergences of integrated information at different positions than the original model. Only a situated version of integrated information, which does not dismiss the activity of the environment and its couplings to the system, can correctly measure integrated information even for a model as simple as ours. This suggest that the intrinsic notion of information or the intrinsic perspective of a system cannot dismiss the system's regulation of its coupling with the environment [22]. Thus, ignoring the coupling with the outside of a sytem can have important consequences for the application of integrated information measures in simulated and experimental setups. For example, in [23], different agents are characterised by the integrated information of its neural mechanisms, but ignoring the environment might miss important channels of sensorimotor coordination contributing to the integration of the system. Similarly, attempts to identify the physical substrate of consciousness in brains [24] should take into account situated and embodied aspects of brain activity, or even consider the possibility that (at least at some moments) this substrate can cut across the brain–body–world divisions, rather than being confined inside the brain [25].

In other experiments, we compared the differences between the values of mechanism-level and system-level integration (φ and Φ) in a homogeneous system with one region, finding that some normalisation constants are required to compare Φ of systems with different size. We also found that Φ also diverges at the critical point, and it does faster than φ, due to the second partition applied and the accumulation of the different mechanisms of the system. Although here we compute Φ for a very simple system, we suggest that the introduction of this second level of analysis should be better justified. In that sense, recent work explores very small networks showing how the compositional definition of measures like Φ can yield very different results than the non-compositional mechanism-level measures φ [26]. Further work could try to better characterise the difference between the two levels in systems with analytically tractable properties like the Ising systems with multiple regions presented here.

Finally, we compared the diverging tendencies of two coupled subregions, showing that the delimitation of integrated information might change with size as the integration of some regions diverges faster than others. This is specially relevant as IIT 3.0 gives a prominent relevance to the areas of the brain with maximal integration (the "neural substrate of consciousness"). If integration takes the form of diverging values of φ around certain classes of critical points, or regions (see [14]), then the neural substrate supporting maximal integration should be characterised by how fast integration diverges with size, and not by the value of integration yielded by simplified models (e.g., by coarse-graining observed time series), which can be potentially misleading.

These results exploring homogeneous kinetic Ising models show that the calculation of integrated information presents important challenges even in simple models. This work serves to demonstrate that the measure is very susceptible to design assumptions and that the behaviour of the measure changes drastically because of this. In this scenario, we show how the connection between the theory (IIT 3.0), a theoretical understanding of complex dynamical systems (critical phase transitions), and the study of simplified models exemplifying known phenomena (homogeneous Ising models) offers a path to systematically study the implications of these assumptions. Our results compel researchers interested in IIT and related indices of complexity to apply such measures under careful examination of their design assumptions. Rather than applying the measure off-the-shelf, ideally researchers should be aware of the assumptions behind the measure and how it applies to each situation. In this way, theory can go in hand with cautious experimental applications, avoiding potentially misleading interpretations and ensuring that they are used to improve our understanding of biological and neural phenomena.

Funding: Miguel Aguilera was supported by the UPV/EHU post-doctoral training program ESPDOC17/17 and supported by project TIN2016-80347-R funded by the Spanish Ministry of Economy and Competitiveness and project IT1228-19 funded by Basque Government.

Acknowledgments: The author would like to thank Ezequiel Di Paolo for constructive criticism of the manuscript.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. List of Assumptions and Experiments

We show here a summary of the different assumptions made by IIT 3.0 and different sections of the manuscript. The different assumptions are described in Table A1. Moreover, in Table A2, we indicate which assumptions are considered by IIT 3.0 and the figures corresponding to different experiments in the article.

Note that due to computational cost of some of the measures, different experiments use different system sizes, and consequently they explore different values of couplings, as in larger systems peaks of values related with critical divergences take place closer to the critical point.

Table A1. Description of the different assumptions considered in the article.

Assumptions	Description
Homogeneous connectivity	In order to simplify the computation of probability distributions and the MIP in the thermodynamic limit, we assume that the system is divided into a number of homogeneous regions. All units within a region share the same inter/intra-region coupling values.
Equal mechanism and purview	To simplify calculations, we assume that the purview of the system is equal to the mechanism. In contrast, IIT 3.0 selects the purview that yields maximum integration φ.
MIP cuts either a single node or entire regions	In the thermodynamic limit, when all couplings are positive, the MIP of a homogeneous Ising model is either a partition that cuts a single node from the mechanism or one that separates an entire region (see Appendix B.3). We assume that the same applies to finite systems.
Initial noise injection	When transition probability matrices describe several updates, IIT 3.0 assumes that partitions only inject noise in the initial state.
Continuous noise injection	In contrast with IIT 3.0, in some cases we assume that partitions inject noise at every update of the system.
Independent prior	In order to compute the cause repertoire of a mechanism, IIT 3.0 assumes a uniform prior distribution to apply Bayes rule (Equation (17))
Stationary prior	Alternative, in some cases we assume a stationary prior to compute cause repertoires of a mechanism (Section 3.2).
Wasserstein distance	In IIT 3.0, distances between distributions are computed using the Wasserstein distance.
Kullback–Leibler divergence	Many other alternative measures of integration (including previous versions of IIT) are based on the Kullback–Leibler divergence.

Table A2. List of assumptions considered by IIT 3.0 and by the results of the different experiments in the article.

Assumptions & Experiments	IIT 3.0	Figure 3A	Figure 3B	Figure 4A	Figure 4B	Figure 5A	Figure 5B,C	Figures 6A and 8	Figure 6B,C	Figure 7A,B
Homogeneous connectivity		✓	✓	✓	✓	✓	✓	✓	✓	✓
Equal mechanism and purview		✓	✓	✓	✓	✓	✓	✓	✓	✓
MIP is a single node or entire regions		✓	✓	✓	✓	✓	✓	✓	✓	✓
Initial noise injection	✓	✓								
Continuous noise injection			✓	✓	✓	✓	✓	✓	✓	✓
Independent prior	✓			✓						
Stationary prior					✓	✓	✓	✓	✓	✓
Wasserstein distance	✓	✓	✓	✓	✓	✓		✓	✓	✓
KL divergence							✓			
Environment decoupling	✓								✓	
Environment coupling								✓		✓

Appendix B. Wasserstein Distance

The Wasserstein distance (or Earth Mover's distance), $D_W(P(\mathbf{s}), P(\mathbf{s}))$, is defined as the minimum "cost of transportation" that arises when transforming one probability distribution $P(\mathbf{s})$ into another $Q(\mathbf{s})$. This cost is defined as the amount of "mass" that has to be moved from each state in $P(\mathbf{s})$ to another in $Q(\mathbf{s})$, defined by the matrix $\mathbf{W}$, and the distance this mass has to be moved, which is defined by the distance $d(\mathbf{s}_I, \mathbf{s}_J)$ this mass has to be transported, defined as the Hamming distance between $\mathbf{s}_I$ and $\mathbf{s}_J$, which counts the number of places by which two strings differ. Thus, the Wasserstein distance is defined as

$$D_W(P(\mathbf{s}), Q(\mathbf{s})) = \min_{\mathbf{W} \in \mathcal{W}(P,Q)} \sum_{\mathbf{s}_I, \mathbf{s}_J} d(\mathbf{s}_I, \mathbf{s}_J) W_{IJ}, \tag{A1}$$

where the indices I, J cover the set of possible states of the array $\mathbf{s}$, and $\mathcal{W}(P, Q)$ represents the set of transfer matrices that meet the conditions $\sum_J W_{IJ} = P(\mathbf{s})$, $\sum_I W_{IJ} = Q(\mathbf{s})$, $\sum_{IJ} W_{IJ} = 1$

Appendix B.1. Finite Size

For networks with homogeneous regions of finite size, the Wasserstein distance can be computed directly from the aggregate variables $\mathbf{S}$. This is justified as follows.

For each value of $\mathbf{S}_K$, there will be a set of corresponding values $\{\mathbf{s}_I\}_{\Sigma(\mathbf{s}_I)=\mathbf{S}_K}$, where $\Sigma(\mathbf{s}_I)$ is the transformation that obtains the aggregate values of each region such that for each region $\mathcal{U}$ we have that $S_{\mathcal{U}}(t) = \sum_{i \in \mathcal{U}} (1 + s_i(t))/2$.

We note that if $\Sigma(\mathbf{s}_I) = \Sigma(\mathbf{s}_J) = \mathbf{S}_K$, this implies in homogeneous systems that probabilities of those states are identical: $P(\mathbf{s}_I) = P(\mathbf{s}_J)$. For a pair of values $\mathbf{S}_K, \mathbf{S}_L$ such that $\Sigma(\mathbf{s}_I) = \mathbf{S}_K, \Sigma(\mathbf{s}_J) = \mathbf{S}_L$, we consider the case in which we transport a set of identical probability distributions $\{P(\mathbf{s}_I)\}_{\Sigma(\mathbf{s}_I)=\mathbf{S}_K}$ into $\{P(\mathbf{s}_J)\}_{\Sigma(\mathbf{s}_J)=\mathbf{S}_L}$. As there is the same amount of mass in all sources and destinations in the transportation cost, there is always a transport scheme such that

$$W_{IJ} = \begin{cases} 0, & \text{if } d(\mathbf{s}_I, \mathbf{s}_J) > \min\limits_{\substack{\mathbf{s}_{I'}, \Sigma(\mathbf{s}_{I'})=\mathbf{S}_K \\ \mathbf{s}_{J'}, \Sigma(\mathbf{s}_{J'})=\mathbf{S}_L}} d(\mathbf{s}_{I'}, \mathbf{s}_{J'}) \\ W^*_{IJ}, & \text{otherwise} \end{cases}. \tag{A2}$$

Moreover,

$$\min_{\substack{\mathbf{s}_{I'},\Sigma(\mathbf{s}_{I'})=\mathbf{S}_K \\ \mathbf{s}_{J'},\Sigma(\mathbf{s}_{J'})=\mathbf{S}_L}} d(\mathbf{s}_{I'},\mathbf{s}_{J'}) = d(\mathbf{S}_K,\mathbf{S}_L) = \sum_{\mathcal{U}} |S_{K,\mathcal{U}} - S_{L,\mathcal{U}}| \tag{A3}$$

In this case, we can rewrite Equation (A1) as

$$D_W(P(\mathbf{s}),Q(\mathbf{s})) = D_W(P(\mathbf{S}),Q(\mathbf{S})) = \min_{\mathbf{W}\in\mathcal{W}(P(\mathbf{S}),Q(\mathbf{S}))} \sum_{\mathbf{S}_K,\mathbf{S}_L} d(\mathbf{S}_K,\mathbf{S}_L) W_{KL}. \tag{A4}$$

Appendix B.2. Infinite Size

In the infinite-size kinetic Ising model with homogeneous connectivity, probability distributions are the product of an array of independent distributions $P(s_i(t\pm\tau)|\mathbf{s}(t)) = \frac{1+s_i(t\pm\tau)m_i(t\pm\tau)}{2}$. Thus, the cost of transport can be defined as the sum of individual costs of the independent distributions:

$$D_W(P(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))||P_{\text{cut}}(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))) = \frac{1}{2}\sum_i \left| m_i(t\pm\tau) - m_i^{\text{cut}}(t\pm\tau) \right|. \tag{A5}$$

The mean activation rate of a region can be interpreted as a function of couplings $\mathbf{J}$, $m_i(t\pm\tau \mid \mathbf{J})$. We describe mean rate of a region under partition cut as $m_i^{\text{cut}(t\pm\tau)} = m_i(t\pm\tau|\mathbf{J}+\mathbf{dJ}))$. Where $dJ_{ij} = J_{ij}$ if $J_{ij} \in \mathcal{J}_{\text{cut}}$; otherwise, $dJ_{ij} = 0$.

We assume a homogeneous system with a number of regions, with $J_{ij} = \frac{1}{N} J_{\mathcal{U},\mathcal{V}}$ and the partition affects a small number of connections (see Appendix B.3), in the thermodynamic limit the mean rate can be described as the first order term of a Taylor expansion at $\mathbf{dJ} = 0$:

$$m_i(t\pm\tau|\mathbf{J}+\mathbf{dJ})) = m_i(t\pm\tau|\mathbf{J})) + \sum_{J_{ij}\in\mathcal{J}_{\text{cut}}} \frac{\partial m_i(t\pm\tau))}{\partial J_{ij}} J_{ij}, \tag{A6}$$

then the Wasserstein distance is

$$D_W(P(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))||P_{\text{cut}}(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))) = \frac{1}{2}\sum_i \left| \sum_{J_{kl}\in\mathcal{J}_{\text{cut}}} \frac{\partial m_i(t\pm\tau))}{\partial J_{kl}} J_{kl} \right|. \tag{A7}$$

We assume that the system is divided into homogeneous regions, for $k \in \mathcal{U}, l \in \mathcal{V}$, and $J_{kl} = \frac{1}{N} J_{\mathcal{U}\mathcal{V}}$.

$$D_W(P(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))||P_{\text{cut}}(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))) = \frac{1}{2}\sum_i \left| \sum_{\mathcal{U}\mathcal{V}} \sum_{\substack{k\in\mathcal{U},l\in\mathcal{V} \\ J_{kl}\in\mathcal{J}_{\text{cut}}}} \frac{\partial m_i(t\pm\tau))}{\partial J_{kl}} J_{kl} \right| \tag{A8}$$

$$= \frac{1}{2}\sum_{\mathcal{W}} N_{\mathcal{W}} \left| \sum_{\mathcal{U}\mathcal{V}} N_{\text{cut}}[\mathcal{U},\mathcal{V}] \frac{\partial m_{\mathcal{W}}(t\pm\tau)}{\partial J_{\mathcal{U}\mathcal{V}}} \frac{J_{\mathcal{U}\mathcal{V}}}{N} \right|, \tag{A9}$$

where $N_{\text{cut}}[\mathcal{U},\mathcal{V}]$ is the number of connections from region $\mathcal{U}$ to region $\mathcal{V}$ that is affected by the partition cut. as it is shown in [14]. Note that the terms $\frac{\partial m_i(t\pm\tau))}{\partial J_{kl}} J_{kl}$ are always equal when $i \neq k$, and the case of $i = k$ can is neglected in the thermodynamic limit. Also note that integrated information for infinite size in the homogeneous kinetic Ising model with one region would be equivalent to the magnetic susceptibility of the system [13].

Appendix B.3. Minimum Information Partition in the Thermodynamic Limit

In the homogeneous kinetic Ising model, any conditional distribution can be computed as a product of independent distributions, by recursively computing Equation (13). When computing

integrated information with infinite sizes using Equation (A9), if all couplings are positive, the sign of all $\frac{\partial m_{\mathcal{W}}(t\pm\tau)}{\partial J_{\mathcal{U}\mathcal{V}}}$ is always the same. Then the MIP is the partition that minimises

$$D_W(P(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))||P_{\text{cut}}(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))) = \frac{1}{2}\sum_{\mathcal{U}\mathcal{V}} N_{\text{cut}}[\mathcal{U},\mathcal{V}]F[\mathcal{U},\mathcal{V}], \tag{A10}$$

$$F[\mathcal{U},\mathcal{V}] = \sum_{\mathcal{W}} N_{\mathcal{W}}\left|\frac{\partial m_{\mathcal{W}}(t\pm\tau)}{\partial J_{\mathcal{U}\mathcal{V}}}\right|\frac{J_{\mathcal{U}\mathcal{V}}}{N}, \tag{A11}$$

as $F[\mathcal{U},\mathcal{V}] > 0$.

We can describe the number of cut partitions as $N_{\text{cut}}[\mathcal{U},\mathcal{V}] = N_{\mathcal{U}}N_{\mathcal{V}}(f^c_{\mathcal{U}} + f^f_{\mathcal{V}} - 2f^c_{\mathcal{U}}f^f_{\mathcal{V}})$, where $f^c_{\mathcal{U}}$ are the fraction of units of region $\mathcal{V}$ cut by the partition in the present state, and $f^f_{\mathcal{V}}$ are the fraction of units of region $\mathcal{U}$ cut by the partition in the future (or past) state. The only constraints are that $f_{\mathcal{U}} = \frac{n_{\mathcal{U}}}{N_{\mathcal{U}}}, n_{\mathcal{U}} \in \mathbb{Z}$, with $0 < n_{\mathcal{U}} < N_{\mathcal{U}}$, $\max(\sum_{\mathcal{U}} n^c_{\mathcal{U}}, \sum_{\mathcal{U}} n^f_{\mathcal{U}} > 0$ and $\min(\sum_{\mathcal{U}} n^c_{\mathcal{U}}, \sum_{\mathcal{U}} n^f_{\mathcal{U}}) < N$.

Take, for example, a specific region $\mathcal{U}'$, we can decompose the distance function in

$$D_W = N_{\mathcal{U}'}N_{\mathcal{V}}(\sum_{\mathcal{V}} f^f_{\mathcal{V}}F[\mathcal{U}',\mathcal{V}] + f^c_{\mathcal{U}'}\sum_{\mathcal{V}}(1-2f^f_{\mathcal{V}})F[\mathcal{U}',\mathcal{V}]) + \sum_{\mathcal{U}\neq\mathcal{U}',\mathcal{V}} N_{\mathcal{U}}N_{\mathcal{V}}(f^c_{\mathcal{U}} + f^f_{\mathcal{V}} - 2f^c_{\mathcal{U}}f^f_{\mathcal{V}})F[\mathcal{U},\mathcal{V}]. \tag{A12}$$

If we only consider changes in the value of $f^c_{\mathcal{U}'}$, the first term is minimised by

$$f^c_{\mathcal{U}'} = \begin{cases} 0, & \text{if } \sum_{\mathcal{V}}(1-2f^f_{\mathcal{V}})F[\mathcal{U}',\mathcal{V}] > 0. \\ 1, & \text{otherwise.} \end{cases} \tag{A13}$$

As well, the function grows monotonically with $f^c_{\mathcal{U}'}$.

Repeating this observation for every possible $f^c_{\mathcal{U}'}, f^f_{\mathcal{V}'}$, if $0 < f < 1$, the MIP should be a partition in which all values of f are either 0 or 1. A trivial solution is one where all $f = 0$ or all $f = 1$, but this case violates one of the last two constraints above. The closest possible solution that complies with the minimum is one in which one $f^c_{\mathcal{U}'} = \frac{1}{N^c_{\mathcal{U}'}}$ or one $f_{\mathcal{V}'} = \frac{1}{N_{\mathcal{V}'}}$, where for the rest of values $f^c_{\mathcal{U}} = 0, f_{\mathcal{V}} = 0$.

That is, the space of possible partitions that could constitute the MIP is constrained to (a) partitions that isolate one single unit in the repertoire of current or future (or past) states or (b) one partition in which all elements of a region in the current or the future (or past) states belong to the same partition.

In case (a), if the unit isolated by the partition belongs to region $\mathcal{U}$ in the current state, the distance is

$$D_W(P(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))||P_{\text{cut}}(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))) = \frac{1}{2}\sum_{\mathcal{V}} N_{\mathcal{V}}F[\mathcal{U},\mathcal{V}], \tag{A14}$$

whereas when the isolated unit belongs to the future (or past) state, the mean rate of the isolated unit i will be $m^{\text{cut}}_i(t\pm\tau) = 0$. Thus, according to Equation (A7), the distance is

$$D_W(P(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))||P_{\text{cut}}(\mathbf{s}(t\pm\tau)|\mathbf{s}(t))) = \frac{1}{2}\left|m_{\mathcal{U}}(t\pm\tau)\right| + \frac{1}{2}(1-\frac{1}{N_{\mathcal{U}}})\sum_{\mathcal{V}} N_{\mathcal{V}}F[\mathcal{U},\mathcal{V}]. \tag{A15}$$

Thus, in the thermodynamic limit, in case (a) the isolated unit will always belong to the present state, and never the purview. An equivalent argument applies to the cause repertoire in a stationary prior is assumed.

Appendix B.4. Minimum Information Partition in the Thermodynamic Limit for Computing Φ

The approximation of the MIP of Φ in the thermodynamic limit of a kinetic Ising model with an homogeneous region with coupling $J_{ij} = \frac{J}{N}$ can be computed by calculating the first term of the Taylor expansion of $D_W^*(C(J), C_{\text{cut}}(J + dJ))$ (from Equation (5)) around $J = 0$:

$$D_W^*(C(J), C_{\text{cut}}(J + dJ)) = N_{\text{cut}} \frac{\partial D_W^*(C(J), C_{\text{cut}}(J + dJ))}{\partial J}\Big|_{dJ=0} \frac{J}{N}, \tag{A16}$$

as in previous cases, if the approximation accurate (the partition is small) then $\frac{\partial D_W^*(C(J), C_{\text{cut}}(J+dJ))}{\partial J}$ should be positive and the MIP will be a partition that cuts one single node, i.e., $N_{\text{cut}} = 1$.

Appendix C. Kullback–Leibler Divergence

Many versions of φ use the Kullback–Leibler divergence as an alternative distance measure to the Wasserstein distance.

The Kullback–Leibler divergence is defined as

$$D_{KL}(P(\mathbf{s})||Q(\mathbf{s})) = \sum_{\mathbf{s}} P(\mathbf{s}) \log \frac{P(\mathbf{s})}{Q(\mathbf{s})}. \tag{A17}$$

Appendix C.1. Finite Size

In the finite size, the Kullback–Leibler divergence $D_{KL}(P(\mathbf{s})||Q(\mathbf{s}))$ is equivalent to the divergence of the aggregate variables $D_{KL}(P(\mathbf{S})||Q(\mathbf{S}))$. This can be shown as follows. The probability $P(\mathbf{S})$ is equal to the sum of all probabilities $P(\Sigma(\mathbf{s}))$ such that $\Sigma(\mathbf{s}) = \mathbf{S}$, having that

$$P(\mathbf{S}) = \sum_{\Sigma(\mathbf{s})=\mathbf{S}} P(\Sigma(\mathbf{s})) = P(\Sigma(\mathbf{s})) \sum_{\mathcal{U}} \binom{N_{\mathcal{U}}}{S_{\mathcal{U}}} \tag{A18}$$

then we have that

$$D_{KL}(P(\mathbf{s})||Q(\mathbf{s})) = \sum_{\mathbf{s}} P(\mathbf{s}) \log \frac{P(\mathbf{s})}{Q(\mathbf{s})} = \sum_{\mathbf{s}} P(\mathbf{s}) \log \frac{P(\mathbf{S})}{Q(\mathbf{S})} = D_{KL}(P(\mathbf{S})||Q(\mathbf{S})) \tag{A19}$$

Appendix C.2. Infinite Size

In the infinite-size kinetic Ising model with homogeneous connectivity, as the state of the units of the system is independent, the Kullback–Leibler divergence can be defined as the sum of individual divergences. As $P(s_i(t \pm \tau)|\mathbf{s}(t)) = \frac{1+s_i(t\pm\tau)m_i(t\pm\tau)}{2}$, we have that

$$\begin{aligned} &D_{KL}(P(\mathbf{s}(t \pm \tau)|\mathbf{s}(t))||P_{\text{cut}}(\mathbf{s}(t \pm \tau)|\mathbf{s}(t))) \\ &\quad = -\sum_i \sum_{s_i} P(s_i(t \pm \tau)|\mathbf{s}(t)) \log \left(1 + \frac{(m_i^{\text{cut}}(t \pm \tau) - m_i(t \pm \tau))s_i}{1 + m_i(t \pm \tau)s_i}\right). \end{aligned} \tag{A20}$$

In the thermodynamic limit, if a partition cuts a small number of connections $m_i^{\text{cut}}(t \pm \tau) - m_i(t \pm \tau)$ is small and we can approximate the value of the distance as a Taylor expansion of order n:

$$\begin{aligned} &D_{KL}(P(\mathbf{s}(t \pm \tau)|\mathbf{s}(t))||P_{\text{cut}}(\mathbf{s}(t \pm \tau)|\mathbf{s}(t))) \\ &\quad = \frac{1}{2} \sum_i \sum_{s_i} \frac{1}{2} \left(\frac{1}{1 + m_i(t \pm \tau)s_i} \sum_{J_{kl}, J_{mn} \in \mathcal{J}_{\text{cut}}} \frac{\partial^2 m_i(t \pm \tau)}{\partial J_{kl} \partial J_{mn}} J_{kl} J_{mn} \right) \\ &\quad = \frac{1}{2} \sum_i \left(\frac{1}{1 - m_i^2(t \pm \tau)} \sum_{J_{kl} J_{mn} \in \mathcal{J}_{\text{cut}}} \frac{\partial^2 m_i(t \pm \tau)}{\partial J_{kl} \partial J_{mn}} \frac{J_{kl} J_{mn}}{N^2} \right). \end{aligned} \tag{A21}$$

We assume that the system is divided into homogeneous regions, for $k \in \mathcal{U}, l \in \mathcal{V}$ and $J_{kl} = \frac{1}{N} J_{\mathcal{UV}}$.

$$\begin{aligned} &D_{KL}(P(\mathbf{s}(t \pm \tau)|\mathbf{s}(t))||P_{\text{cut}}(\mathbf{s}(t \pm \tau)|\mathbf{s}(t))) \\ &\quad = \frac{1}{2} \sum_i \frac{1}{1 - m_i^2(t \pm \tau)} \sum_{\mathcal{UV}} \sum_{\substack{k \in \mathcal{U}, l \in \mathcal{V} \\ J_{kl} \in \mathcal{J}_{\text{cut}}}} \sum_{\mathcal{ST}} \sum_{\substack{m \in \mathcal{S}, n \in \mathcal{T} \\ J_{mn} \in \mathcal{J}_{cut}}} \frac{\partial^2 m_{\mathcal{W}}(t \pm \tau)}{\partial J_{\mathcal{UV}} \partial J_{\mathcal{ST}}} J_{\mathcal{UV}} J_{\mathcal{ST}} \\ &\quad = \frac{1}{2} \sum_{\mathcal{W}} \frac{N_{\mathcal{W}}}{1 - m_{\mathcal{W}}^2(t \pm \tau)} \sum_{\mathcal{UVST}} N_{\text{cut}}[\mathcal{U}, \mathcal{V}] N_{\text{cut}}[\mathcal{S}, \mathcal{T}] \frac{\partial^2 m_{\mathcal{W}}(t \pm \tau)}{\partial J_{\mathcal{UV}} \partial J_{\mathcal{ST}}} J_{\mathcal{UV}} J_{\mathcal{ST}}. \end{aligned} \tag{A22}$$

where $N_{\text{cut}}[\mathcal{U}, \mathcal{V}]$ is the number of connections from region $\mathcal{U}$ to region $\mathcal{V}$ that is affected by the partition cut. as it is shown in [14]. Note that the terms $\frac{\partial m_i(t \pm \tau))}{\partial J_{kl}} J_{kl}$ are always equal when $i \neq k$, and the case of $i = k$ can is neglected in the thermodynamic limit.

Appendix C.3. Minimum Information Partition in the Thermodynamic Limit for the Kullback–Leibler Divergence

The approximation of the MIP in the thermodynamic limit is harder in the case of the Kullback–Leibler divergence due to the quadratic terms. However, in this article we only compute φ using the Kullback–Leibler divergence in the case of one region with couplings J. In this case, Equation (A22) becomes

$$\begin{aligned} &D_{KL}(P(\mathbf{s}(t \pm \tau)|\mathbf{s}(t))||P_{\text{cut}}(\mathbf{s}(t \pm \tau)|\mathbf{s}(t))) \\ &\quad = \frac{1}{2} \frac{N}{1 - m^2(t \pm \tau)} N_{\text{cut}}^2 \frac{\partial^2 m(t \pm \tau)}{\partial J^2} \frac{J}{N}^2, \end{aligned} \tag{A23}$$

where N_{cut} is the number of connections cut by the partition. In this case, as $\frac{\partial^2 m(t \pm \tau)}{\partial J^2} J^2$ is always positive for positive couplings, we have that the MIP is just a partition that cuts one single node.

References

1. 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. [CrossRef] [PubMed]
2. Marshall, W.; Kim, H.; Walker, S.I.; Tononi, G.; Albantakis, L. How causal analysis can reveal autonomy in models of biological systems. *Phil. Trans. R. Soc. A* **2017**, *375*, 20160358. doi:10.1098/rsta.2016.0358. [CrossRef] [PubMed]
3. Miyahara, K.; Witkowski, O. The integrated structure of consciousness: Phenomenal content, subjective attitude, and noetic complex. *Phenom. Cogn. Sci.* **2019**, *18*, 731–758. [CrossRef]
4. Cerullo, M.A. The Problem with Phi: A Critique of Integrated Information Theory. *PLoS Comput. Biol.* **2015**, *11*. doi:10.1371/journal.pcbi.1004286. [CrossRef]
5. Barrett, A.B.; Mediano, P.A. The Phi measure of integrated information is not well-defined for general physical systems. *J. Conscious. Stud.* **2019**, *26*, 11–20.
6. Mediano, P.A.M.; Seth, A.K.; Barrett, A.B. Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. *Entropy* **2019**, *21*, 17. doi:10.3390/e21010017. [CrossRef]
7. Salinas, S.R.A. The Ising Model. In *Introduction to Statistical Physics*; Salinas, S.R.A., Ed.; Graduate Texts in Contemporary Physics; Springer: New York, NY, USA, 2001; pp. 257–276. doi:10.1007/978-1-4757-3508-6_13. [CrossRef]
8. Salinas, S.R.A. Scaling Theories and the Renormalization Group. In *Introduction to Statistical Physics*; Springer: New York, NY, USA, 2001; pp. 277–304.
9. Beggs, J.M. The criticality hypothesis: How local cortical networks might optimize information processing. *Philos. Trans. R. Soc. A* **2007**, *366*, 329–343. [CrossRef]
10. Barrett, A.B.; Seth, A.K. Practical Measures of Integrated Information for Time-Series Data. *PLoS Comput. Biol.* **2011**, *7*, e1001052. doi:10.1371/journal.pcbi.1001052. [CrossRef]
11. Oizumi, M.; Amari, S.i.; Yanagawa, T.; Fujii, N.; Tsuchiya, N. Measuring Integrated Information from the Decoding Perspective. *PLoS Comput. Biol.* **2016**, *12*, e1004654. doi:10.1371/journal.pcbi.1004654. [CrossRef]

12. Hoel, E.P.; Albantakis, L.; Marshall, W.; Tononi, G. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neurosci. Conscious.* **2016**, *2016*. doi:10.1093/nc/niw012. [CrossRef]
13. Kochmański, M.; Paszkiewicz, T.; Wolski, S. Curie-Weiss magnet: A simple model of phase transition. *Eur. J. Phys.* **2013**, *34*, 1555–1573. doi:10.1088/0143-0807/34/6/1555. [CrossRef]
14. Aguilera, M.; Di Paolo, E. Integrated information in the thermodynamic limit. *Neural Netw.* **2019**. doi:10.1016/j.neunet.2019.03.001. [CrossRef] [PubMed]
15. Pressé, S.; Ghosh, K.; Lee, J.; Dill, K.A. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.* **2013**, *85*, 1115–1141. doi:10.1103/RevModPhys.85.1115. [CrossRef]
16. Tegmark, M. Consciousness as a state of matter. *Chaos Soliton. Fract.* **2015**, *76*, 238–270. [CrossRef]
17. Marshall, W.; Albantakis, L.; Tononi, G. Black-boxing and cause-effect power. *PLoS Comput. Biol.* **2018**, *14*, e1006114. doi:10.1371/journal.pcbi.1006114. [CrossRef]
18. Chiel, H.J.; Beer, R.D. The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. *Trends Neurosci.* **1997**, *20*, 553–557. [CrossRef]
19. Clark, A. The Dynamical Challenge. *Cogn. Sci.* **1997**, *21*, 461–481. doi:10.1207/s15516709cog2104_3. [CrossRef]
20. Aguilera, M.; Bedia, M.G.; Santos, B.A.; Barandiaran, X.E. The situated HKB model: How sensorimotor spatial coupling can alter oscillatory brain dynamics. *Front. Comput. Neurosci.* **2013**, *7*. [CrossRef]
21. Di Paolo, E.; Buhrmann, T.; Barandiaran, X. *Sensorimotor Life: An Enactive Proposal*; Oxford University Press: Oxford, UK, 2017.
22. Di Paolo, E.A. Autopoiesis, Adaptivity, Teleology, Agency. *Phenomenol. Cogn. Sci.* **2005**, *4*, 429–452. doi:10.1007/s11097-005-9002-y. [CrossRef]
23. Albantakis, L.; Hintze, A.; Koch, C.; Adami, C.; Tononi, G. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLOS Comput. Biol.* **2014**, *10*, e1003966. [CrossRef]
24. Tononi, G.; Boly, M.; Massimini, M.; Koch, C. Integrated information theory: From consciousness to its physical substrate. *Nat. Rev. Neurosci.* **2016**, *17*, 450–461. doi:10.1038/nrn.2016.44. [CrossRef] [PubMed]
25. Thompson, E.; Varela, F.J. Radical embodiment: Neural dynamics and consciousness. *Trends Cogn. Sci.* **2001**, *5*, 418–425. doi:10.1016/S1364-6613(00)01750-2. [CrossRef]
26. Albantakis, L.; Tononi, G. Causal Composition: Structural Differences among Dynamically Equivalent Systems. *Entropy* **2019**, *21*, 989. doi:10.3390/e21100989. [CrossRef]