Integrated information in the thermodynamic limit

Miguel Aguilera [a,b,*], Ezequiel A. Di Paolo [a,c,d]

[a] *IAS-Research Center for Life, Mind, and Society, University of the Basque Country, Donostia, Spain*
[b] *ISAAC Lab, Aragón Institute of Engineering Research, University of Zaragoza, Zaragoza, Spain*
[c] *Ikerbasque, Basque Foundation for Science, Bizkaia, Spain*
[d] *Centre for Computational Neuroscience and Robotics, Department of Informatics, University of Sussex, Brighton, UK*

A B S T R A C T

The capacity to integrate information is a prominent feature of biological, neural, and cognitive processes. Integrated Information Theory (IIT) provides mathematical tools for quantifying the level of integration in a system, but its computational cost generally precludes applications beyond relatively small models. In consequence, it is not yet well understood how integration scales up with the size of a system or with different temporal scales of activity, nor how a system maintains integration as it interacts with its environment. After revising some assumptions of the theory, we show for the first time how modified measures of information integration scale when a neural network becomes very large. Using kinetic Ising models and mean-field approximations, we show that information integration diverges in the thermodynamic limit at certain critical points. Moreover, by comparing different divergent tendencies of blocks that make up a system at these critical points, we can use information integration to delimit the boundary between an integrated unit and its environment. Finally, we present a model that adaptively maintains its integration despite changes in its environment by generating a critical surface where its integrity is preserved. We argue that the exploration of integrated information for these limit cases helps in addressing a variety of poorly understood questions about the organization of biological, neural, and cognitive systems.

1. Introduction

Cognition emerges from the distributed activity of many neural, bodily, and environmental processes. The problem of large-scale integration of neural processes is crucial for understanding how unified cognitive and behavioural states arise from the coordination of these distributed sources of activity. Evidence (Bassett & Gazzaniga, 2011; Pessoa, 2014) suggests that this integration process is non-decomposable: we cannot understand it in terms of modular components or timescales of activity in a neural system nor can we decouple neural activity from the external environment (Aguilera, Bedia, Santos, & Barandiaran, 2013). The different components and scales of cognitive processes are deeply intertwined. Yet, functional components are still able to maintain their differentiated characteristics and generate complex adaptive patterns of behaviour.

How can such an integrated, complex organization emerge and be maintained? One of the most attractive theories we currently have is that neural activity is coordinated into a coherent yet flexible 'dynamic core' (Tononi & Edelman, 1998; Varela, 1995), which balances opposing tendencies of integration and segregation. The interplay of these opposing tendencies generates information (understood as described by information theory, not in a semantic or intensional sense) that is highly diversified among functional parts of the nervous system, and at the same time unified into a coherent whole, thus displaying highly complex patterns of activity.

Integrated information is defined as the information possessed by a system which is above and beyond the information that is available from the sum of its parts. Information integration was first conceived of as linked to consciousness (Oizumi, Albantakis, & Tononi, 2014; Tononi & Edelman, 1998) but it can also be manifested without awareness (Mudrik, Faivre, & Koch, 2014) and has been used more generally to describe biological autonomy (Marshall, Kim, Walker, Tononi, & Albantakis, 2017). Although the idea of information integration has received interest from different communities in recent years, we are still lacking a full understanding of the principles that underlie this fundamental process. We know little about how integrative forces are deployed temporally or spatially or how they cope with the surrounding environment. We also do not know how information integration scales up with the size of the system. Should we expect larger systems to be more or less easily integrated than smaller ones?

* Corresponding author at: ISAAC Lab, Aragón Institute of Engineering Research, University of Zaragoza, Zaragoza, Spain.
E-mail address: sci@maguilera.net (M. Aguilera).

Integrated information has been formalized from different perspectives; one of the most popular has been developed as a measure connected to consciousness under the name of *integrated information theory* (IIT, Oizumi et al., 2014). In its latest versions, IIT is based on interventionist notions of causality to characterize the causal influences between the components of a system (Marshall et al., 2017; Oizumi et al., 2014). That is, instead of assessing whether a system is unified into a coherent whole by analysing its behaviour in regular conditions, IIT proposes that the forces integrating the behaviour of the system are better captured by observing its behaviour under perturbations.

IIT postulates that any subset of elements of the system is a *mechanism*[1] integrating information if its intrinsic cause–effect power (i.e., its ability to determine past and future states) is irreducible. Irreducibility is measured in terms of integrated information φ, which when larger than 0 indicates that the subset of elements at its current state constrains the past and future states of the system in a way that cannot be decomposed in two or more independent cause–effect sets of relations. That is, φ captures the level of irreducibility of the system, understood in the sense that even the least disrupting bipartition of the system into two disconnected halves (this is called the *minimum information partition*, MIP) would imply a loss of information in the causal power of the system. Aside from computing integrated information at the level of mechanisms, IIT postulates a composite measure Φ, which is computed from the set of all mechanisms (each one defined by a value of φ) computed in the original system and the system under bidirectional partitions. A system with $\Phi > 0$ is described as forming an irreducible unitary whole. Since many subsets of the system may present $\Phi > 0$, the boundaries of the system are defined around the subset with larger Φ. A detailed description of IIT measures is provided in Appendix A.

Despite usefully formalizing intuitive notions, current formulations of IIT present some limitations for studying brain organization. We propose that, in order to extend current uses of IIT to capture important aspects of neural organization, we should re-examine some of the main assumptions behind its conception:

- **Scalability**. A system can present different levels of integration at different spatial and temporal scales (Hoel, Albantakis, Marshall, & Tononi, 2016; Marshall, Albantakis, & Tononi, 2018) and, in general, it is not well understood how integration behaves at different scales. However, analyses of the properties of brain-inspired statistical mechanical models have unveiled that many processes in neural systems take the form of phase transitions occurring in the thermodynamic limit, showing properties that diverge as the size of the system scales up. Here we apply models from statistical mechanics to describe integration in terms of the tendencies of the system near the thermodynamic limit.

- **Temporal deployment**. The latest formulations of IIT (Oizumi et al., 2014) attempt to capture the dynamical nature of neural systems by focusing on the dynamics of causal processes, not taking the stationarity or ergodicity of the system as initial assumptions. Nevertheless, IIT is only measured at a single scale of temporal activity, since it analyses integration in the causal power of a mechanism from one time step to the next. We propose a modification of φ to study integration along different temporal spans, showing that systems at critical points must be evaluated for very long timescales.

- **Non-decomposability**. As we mentioned, empirical evidence points to the non-decomposability of cognitive processes. In its current formulation, IIT considers elements outside the system under analysis as independent sources of noise. Here, we propose instead that the level of integration of a system must be evaluated in the context of the other systems it is coupled to (therefore not assuming that elements in the environment are just sources of statistical noise). This modification allows us to correctly determine the boundary between a system and its environment in the thermodynamic limit.

Taking these aspects into consideration, we introduce a model to study the spatial and temporal scaling properties of integrated information and the non-decomposability of complex systems. To do so, we modify the IIT framework to match our assumptions: extending the temporal span of the measure, taking into account that a mechanism and its environment are not always decomposable, together with some simplifying assumptions derived from the homogeneity of the regions present in our model. A detailed comparison between the original measure and our modified version can be found in Appendix B.

The issues of scalability and non-decomposability are difficult to address, which is in part due to computational complexity that limits the application of current IIT measures to very small systems and short timescales. IIT has been tested in small toy models (e.g., Albantakis, Hintze, Koch, Adami, & Tononi, 2014; Oizumi et al., 2014, although some alternative formulations try to circumvent this problem, see Barrett & Seth, 2011; Oizumi, Amari, Yanagawa, Fujii & Tsuchiya, 2016).

In contrast, our approach, apart from the modifications proposed above, introduces some simplifications and approximations in order to measure integrated information as a system scales to very large sizes. Specifically, in line with recent efforts to extend IIT theory and clarify its assumptions (Kitazono, Kanai, & Oizumi, 2018; Mediano, Seth, & Barrett, 2019; Oizumi, Amari et al., 2016; Oizumi, Tsuchiya & Amari, 2016), we introduce a simple kinetic Ising model of infinite size and quasi-homogeneous connectivity. Ising models have been used in the past to model the behaviour of neural populations such as in Schneidman, Berry, Segev, and Bialek (2006). And in our case they remain general enough to account for other distributed non-neural processes. We show that, when the size of the model tends to infinity, there exists an exact mean field solution that allows the calculation of the integrated information φ of a system.

We proceed as follows. First, we introduce the kinetic Ising model and a mean field approximation for solving it. Then, we introduce a measure of integrated information and how it can be computed for Ising models of infinite size. Finally, we present the results of our method in three scenarios of increasing complexity depicting how integrated information can be used to characterize an integrated system interacting with an environment:

- In the first scenario, we study the measure in a simple homogeneous model. In the thermodynamic limit, we can describe integrated information as the susceptibility of the system to changes in the direction of the minimum information partition (MIP). Consequently, integrated information diverges when the system is near a critical point.

- The second scenario involves a system coupled to an external environment, showing the system and the system–environment compound both show integrated information diverging near a shared critical point. Nevertheless, depending on the coupling strength, the system and system–environment mechanisms present different speeds of divergence. This allows us to identify the dominant dynamical unit where integration takes place.

---

[1] We use the term *mechanism* in the technical sense described later and not in the specific sense of efficient causality of the mechanical kind. We acknowledge that different forms of causal and enabling relations between processes are possible and relevant, yet we retain the term mechanism in this context to remain coherent with the existing literature.

- Finally, we tune the parameters of a system with internal self-regulation in order to present high integration when interacting with a variety of environments. The system's internal inhibitory interactions generate a critical surface in the direction of the MIP which describes the viable region in which its integration is maintained.

The results presented here represent a first attempt at using integrated information theory to delimit the boundaries of a family of infinite size systems that can be formally solved. The interest of the study is twofold. First, it allows us to check some of the assumptions of IIT and propose some modifications to maintain its consistency in the thermodynamic limit, and to propose a way to adapt IIT measures for very large systems. Second, although the results presented are obtained from relatively simple cases, they offer an opportunity to speculate about how the causal integrative forces of a system (both its internal cohesion and the coupling with its environment) might scale up when a system approaches the thermodynamic limit. This provides an opportunity to address unanswered questions about integrated organization of biological and cognitive systems.

2. Model

We start by describing a general model defining causal interactions between variables. Looking for generality, we use the least structured statistical model (i.e., a maximum caliber model, Pressé, Ghosh, Lee, & Dill, 2013) defining causal correlations between pairs of units from one time step to the next. We study a kinetic Ising model where N binary variables (Ising spins) s_i evolve in discrete time, with synchronous parallel dynamics (Fig. 1.A). Given the configuration of spins at the previous step, $s(t-1) = \{s_1(t-1), \ldots, s_N(t-1)\}$, the spins $s_i(t)$ are independent random variables drawn from the distribution:

$$p(s_i(t)|s(t-1)) = \frac{1}{1 + e^{-2\beta s_i(t)h_i(t)}} \tag{1}$$

where

$$h_i(t) = H_i + \sum_j J_{ij} s_j(t-1) \tag{2}$$

The parameters H_i and J_{ij} 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 $\beta = 1$.

2.1. Mean field kinetic Ising model

We focus on the particular case of a system of infinite size where $H_i = 0$. The system is divided into different regions (from 1 to 3 depending on the example), and the coupling values J_{ij} are positive and homogeneous for each intra- or inter-region connections $J_{ij} = \frac{1}{N_{\mathcal{R}}} J_{\mathcal{S}\mathcal{R}}$, where $\mathcal{R}$ and $\mathcal{S}$ are regions of the system with sizes $N_{\mathcal{R}}, N_{\mathcal{S}}$ and $i \in \mathcal{S}, j \in \mathcal{R}$.

For a system of infinite size (and all regions with also infinite size), the mean field approximation of the input received a unit i belonging to region $\mathcal{S}$ is:

$$h_i(t) = \sum_{\mathcal{R}} J_{\mathcal{S}\mathcal{R}} m_{\mathcal{R}}(t-1),$$
$$m_{\mathcal{R}}(t-1) = \frac{1}{N_{\mathcal{R}}} \sum_{j \in \mathcal{R}} s_j(t-1) \tag{3}$$

where $m_{\mathcal{R}}(t-1)$ is the mean field of region $\mathcal{R}$. Now we can exactly define the update of the mean field variables using Eq. (1) as:

$$m_{\mathcal{S}}(t) = \tanh\left(\sum_{\mathcal{R}} J_{\mathcal{S}\mathcal{R}} m_{\mathcal{R}}(t-1)\right) \tag{4}$$
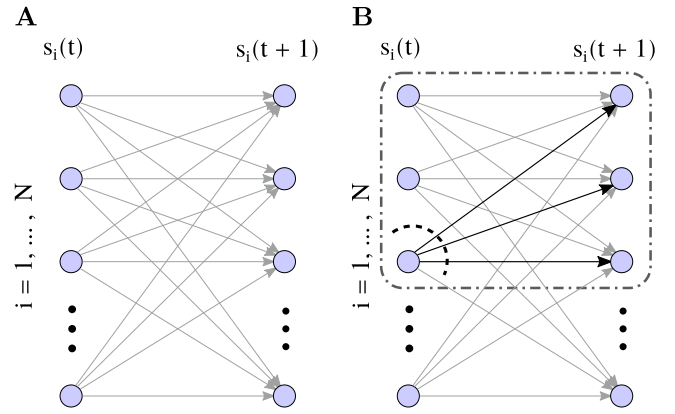

Fig. 1. Kinetic Ising model. A: Description of the infinite size kinetic Ising model. B: Example of a possible partition: the mechanism is delimited by the dot-dashed line and the partition is defined by the dashed line. Partitioned connections (black arrows) are injected with uniform random noise.

2.2. Integrated information φ

We use a simplified version of the integrated effect information described by IIT (Oizumi et al., 2014), implementing some modifications to measure the scaling of integrated information in the thermodynamic limit. In IIT, both causes and effects of a state are taken into account. For simplicity, we consider only the effects of a particular state. Also, although IIT is defined only for the immediate effects after one update of the state of the system, we define integrated information $\varphi(\tau)$ for an arbitrary number of updates of the system. See Appendix B for a list of the differences between IIT and the measure employed here.

Given an initial state $s(\tau_0)$, we define a 'mechanism' $\mathcal{M}$ (following IIT's nomenclature) as a subset of units $\{s_i(\tau_0)\}_{i \in \mathcal{M}}$. The integrated information of mechanism $\mathcal{M}$, $\varphi_{\mathcal{M}}$, is defined as the distance between the behaviour of the original system to a system in which a partition (taken from the set of possible bipartitions) is applied over the units in $\mathcal{M}$. Fig. 1.B depicts an example of a partition. When a partition is applied, the input coming from the partitioned connections of the system is replaced by a random unconstrained noise (binary white noise in the case of an Ising model).

Once the partition is applied, the probability of the state $s(\tau_0 + \tau)$ is computed after τ updates, injecting noise at the partitioned elements during each update. Then, integrated information is defined as the distance D between the conditional probability distributions at $t + \tau$:

$$\varphi_{\mathcal{M}}^{cut}(\tau) = D(p(s(\tau_0+\tau)|s(\tau_0)), p^{cut}(s(\tau_0+\tau)|s(\tau_0))) \tag{5}$$

where $D(p_1, p_2)$ refers to the Wasserstein distance (also known as earth mover's distance) used by IIT to quantify the statistical distance between probability distributions. Here cut specifies the partition applied over the elements of mechanism $\mathcal{M}$, $cut = \{S_1^c, S_2^c, S_1^f, S_2^f\}$, where S_1^c, S_2^c design the blocks of a bipartition of the mechanism at the current state $\{s_i(t)\}_{i \in \mathcal{M}}$, and S_1^f, S_2^f refer to the blocks of a bipartition (not necessarily the same) of the updated state of the units $\{s_i(t+1)\}_{i \in \mathcal{M}}$. Fig. 1.B represents the partition $cut = \{\{s_1(t), s_2(t)\}, \{s_3(t)\}, \{s_1(t+1), s_2(t+1), s_3(t+1)\}, \{\}\}$.

IIT computes integrated information as the value of φ^{cut} under the minimum information partition (MIP), which is the partition of mechanism with the least difference to the original partition (i.e., $\varphi_{\mathcal{M}}^{MIP}(\tau) = \min_{cut} \varphi_M^{cut}(\tau)$). We use $\varphi_{\mathcal{M}}(\tau)$ to denote the minimum information partition integrated information $\varphi_{\mathcal{M}}^{MIP}(\tau)$.

Note that some important modifications have been made. The most important one is that IIT considers the element outside of the mechanism as unconstrained sources of noise. As we show in Fig. B.2, this can radically change the results of integrated information theory, provoking spurious divergences at points other than the critical point. To preserve the consistency of our results, we let elements outside the mechanism operate normally (see Appendix B.3 for details).

2.3. Integrated information in the mean field model

We now show how integrated information can be computed for the Ising model. Thanks to the mean field approximation we can simplify the calculation of the probability distributions of trajectories $p(s(\tau_0 + \tau)|s(\tau_0))$, $p^{cut}(s(\tau_0 + \tau)|s(\tau_0))$ to a Markovian distribution dependent on the mean field at the previous step.

In general, $p(s(\tau_0 + \tau)|s(\tau_0))$ can be computed recursively applying the equation:

$$p(s(\tau_0 + \tau)|p(\tau_0)) =$$
$$\sum_{s(\tau_0+\tau-1)} p(s(\tau_0 + \tau)|s(\tau_0 + \tau - 1))p(s(\tau_0 + \tau - 1)|s(\tau_0)) \quad (6)$$

In the kinetic Ising model of infinite size, the mean fields of the system's regions are deterministic, and instead of computing all possible paths of the system we can just determine the evolution of the mean field using Eq. (4). Moreover, knowing the mean field of each region we can calculate the value of the effective fields $h(\tau_0 + \tau)$ received by each unit using Eq. (3). Also, given the mean field value at a specific point, the posterior probability distribution of each unit is independent. Thus, using the value of $h(\tau_0 + \tau)$ computed evolving from $s(\tau_0)$ we can just take:

$$p(s_i(\tau_0 + \tau)|s(\tau_0)) = p(s_i(\tau_0 + \tau)|h_i(\tau_0 + \tau)) \quad (7)$$

In this case, the calculation of the Wasserstein distance D is drastically simplified, and we can compute φ as the sum of distances between independent binary variables, which is equivalent to computing the difference of their mean values:

$$\varphi_{\mathcal{M}}^{cut}(\tau) = \frac{1}{2} \sum_{\mathcal{R}} N_{\mathcal{R}} |m_{\mathcal{R}}(\tau_0 + \tau) - m_{\mathcal{R}}^{cut}(\tau_0 + \tau)| \quad (8)$$

Once we can calculate φ, we still have the problem of finding the MIP of the system. Luckily, since the connectivity of the system is homogeneous for all nodes in the same region, finding the MIP is equivalent to finding the partition that cuts the lowest number of connections. For infinite size systems where inter-region connections are not zero, the MIP will be one of the possible partitions that isolate just one node of the system. Also, 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 $t + 1$, since partitioning the posterior distribution corresponds to a larger difference between $m_{\mathcal{R}}(\tau_0 + \tau)$ and $m_{\mathcal{R}}^{cut}(\tau_0 + \tau)$. Thus, finding the MIP corresponds to finding which region $\mathcal{R}$ of the system least affects future states when one node of the region is isolated in the partition at time t (e.g., Fig. 1.B).

Finally, we define a function $F_{\mathcal{R}}(m(\tau_0), \tau, \{J_{\mathcal{S},\mathcal{R}}\})$ that recursively applies the update rule in Eq. (4) for τ steps starting from an initial value with a mean field value $m(\tau_0)$, such that $m_{\mathcal{R}}(\tau_0 + \tau) = F_{\mathcal{R}}(m(\tau_0), \tau, J)$. In our mean field approximation, applying the MIP to the quasi-homogeneous system described here is equivalent to just removing one connection[2] between

one or more pairs of regions $\{\mathcal{S}, \mathcal{R}\}_{cut}$, whereas the connections between the rest of regions $\{\mathcal{S}, \mathcal{R}\}_{uncut}$ remain intact. Therefore, the update rule applied by function F to the partitioned system is $F(m(\tau_0), \tau, \{\{J_{\mathcal{S},\mathcal{R}}\}_{uncut}, \{(1 - \frac{1}{N_{\mathcal{R}}})J_{\mathcal{S},\mathcal{R}}\}_{cut}\})$.

Assuming that the number of units per region is equal to $N_{\mathcal{R}} = r_{\mathcal{R}}N$ and $\sum r_{\mathcal{R}} = 1$, we get a simplified expression for the partitioned and unpartitioned terms:

$$F_{\underset{cut}{\mathcal{R}}}(m_0, \tau, x) = F_{\mathcal{R}}(m_0, \tau, \{\{J_{\mathcal{S},\mathcal{R}}\}_{uncut}, \{(1 - \frac{x}{r_{\mathcal{R}}})J_{\mathcal{S},\mathcal{R}}\}_{cut}\})| \quad (9)$$

where $m_0 = m(\tau_0)$ and $x = \frac{1}{N}$ in the partitioned case and $x = 0$ otherwise. Now, computing the unpartitioned and partitioned cases case is equivalent to calculating $F_{\underset{cut}{\mathcal{R}}}(m_0, \tau, x)$ and $F_{\underset{cut}{\mathcal{R}}}(m_0, \tau, \frac{1}{N})$ respectively. Given this, assuming $N \to \infty$ we calculate the final form of φ as a sum of the derivatives of function $F_{\underset{cut}{\mathcal{R}}}(m_0, \tau, x)$:

$$\varphi_{\mathcal{M}}^{cut}(\tau) = \lim_{N \to \infty} \frac{1}{2} \sum_{\mathcal{R}} N_{\mathcal{R}} |F_{\underset{cut}{\mathcal{R}}}(m_0, \tau, 0) - F_{\underset{cut}{\mathcal{R}}}(m_0, \tau, \frac{1}{N})|$$
$$= \frac{1}{2} \sum_{\mathcal{R}} |r_{\mathcal{R}} F'_{\underset{cut}{\mathcal{R}}}(m_0, \tau, 0)| \quad (10)$$

where $F'(m_0, \tau, x) = \frac{dF(m_0, \tau, x)}{dx}$. Note that this defines integrated information in similar terms as the magnetic susceptibility typically used in Ising model to identify critical points, although in this case the mean field of the system is differentiated along the parametrical direction of the MIP.

3. Results

3.1. Integrated information in a homogeneous kinetic Ising model

As an example, we compute numerically the value of $\varphi_{\mathcal{M}_N}(\tau)$ for a homogeneous kinetic Ising model containing just one region (as in Fig. 1.A). The system has only one parameter J describing all connections in the system.

For different values of J, we compute φ for the system starting from a state in the stationary solution. For doing so, we need to know how to compute $F_{cut}(m_0, \tau, x)$, that is, how to compute the mean field of units at a particular time.

First, we numerically compute $F_{cut}(m_0, \tau, x)$ and $\varphi_{\mathcal{M}_N}$ for different values of J for the largest mechanism $\mathcal{M}_N$ of size N, and different values of τ and $m(\tau_0)$ equal to the value at the stationary solution of the system. We estimate the values of the derivative as $F'_{cut}(m_0, \tau, 0) = (F_{cut}(m_0, \tau, dx) - F_{cut}(m_0, \tau, 0))/dx$, using a value $dx = 10^{-10}$. As we observe in Fig. 2.B, the value of $\varphi_{\mathcal{M}_N}(\tau)$ appears to diverge as τ grows.[3]

Similarly, we numerically compute $\varphi_{\mathcal{M}_N}(\tau \to \infty)$ by using the mean field of the model iterating the equation $m(t) = tanh(Jm(t - 1))$ until the difference in the update is smaller than 10^{-15}. In Fig. 2.C we observe that $\varphi_{\mathcal{M}_N}(\tau \to \infty)$ shows an apparent divergence around $J = 1$. Also, we compute the value of $\varphi_{\mathcal{M}_M}(\tau \to \infty)$ for different mechanisms of size M as a fraction of N. As shown in Fig. 2.D, the resulting value of integrated information still diverges but is smaller than the value of $\varphi_{\mathcal{M}_N}(\tau)$ of the whole system, indicating that the system is irreducible.

We can go beyond numerical computations and calculate the analytic value of $\varphi_{\mathcal{M}_N}(\tau \to \infty)$ near the point of divergence by approximating the values of $F_{cut}(m_0, \tau \to \infty, 0)$ around $J = 1$ as the value of m that solves $m = tanh(Jm)$. Note that, more generally, we can compute $F_{cut}(m_0, \tau \to \infty, x)$ by changing the value of the coupling as $J \leftarrow J(1 - x)$.

---

[2] Note that cutting a connection implies injecting uniform noise, which in the mean field approximation is equivalent to substituting the input by a zero mean field or just removing the connection. This is an important approximation that allows us to obtain the main results of the paper, although it will only be valid when the size of the system is infinite and τ is larger than 1.

[3] Note that for larger τ the partition is applied for a longer period of time, and therefore yielding larger integration in some cases.
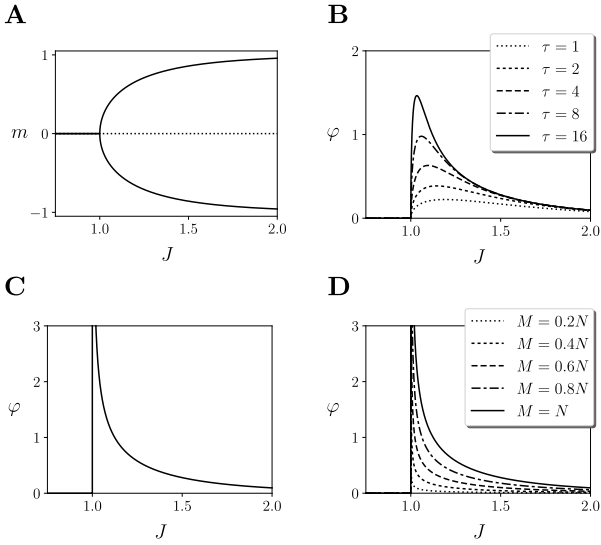
A

B

C

D

Fig. 2. Homogeneous kinetic Ising model. A: Mean activation value in the stationary state. B: Value of $\varphi_{\mathcal{M}_N}(\tau)$ for different temporal spans. C: Value of $\varphi_{\mathcal{M}_N}(\tau \to \infty)$ for an infinite temporal span. D: Value of $\varphi_{\mathcal{M}_M}(\tau \to \infty)$ for different mechanisms of size M and an infinite temporal span.

The system has a trivial solution at $m = 0$. Also, for $J > 1$ the solution at $m = 0$ becomes unstable and a pair of solutions in a pitchfork bifurcation (Fig. 2.A). Although there is no analytic solution of the problem, we can compute the value of m near $J = 1$ by approximating the hyperbolic tangent by its third order Taylor expansion, finding that in the limit $J \to 1^+$ we approximate:

$$F_{cut}(m_0, \tau \to \infty, x) = \pm \sqrt{\frac{3(J(1-x) - 1)}{(J(1-x))^3}}$$

$$\varphi_{\mathcal{M}_N}(\tau \to \infty) = \frac{1}{2}\left| \frac{\sqrt{3}\,(2J - 3)}{2\sqrt{J^3(J-1)}} \right| \tag{11}$$

Thus, we can confirm that the value of integrated information $\varphi_{\mathcal{M}_N}(\tau \to \infty)$ diverges when $J \to 1^+$. This has interesting implications. If a system must maintain a growing level of integration as its size increases, it needs to be poised near a critical point that shows a divergence of the values of φ.

3.2. Integrated information for measuring agent–environment asymmetries

We apply the proposed measure of integrated information to the problem of determining the boundaries of an agent interacting with an environment. One of the central aspects of agency is the existence of agent–environment asymmetries (Barandiaran, Di Paolo, & Rohde, 2009), in which the part of the system corresponding to the agent is able (to some extent) to modulate the terms in which it relates to the surrounding milieu. We test our measure in two simple cases of systems presenting asymmetries in their interaction.

We model a minimal case of agent–environment bidirectional interaction with two regions, where only the region corresponding to the 'agent' has the capacity to self-regulate through recurrent connections (Fig. 3.A). In this case, we have two regions A and E, only A presenting self-connections. The mean field of the system is updated as:

$$m_A(t + 1) = \tanh\left(\frac{1}{2}(J_{AA}m_A(t) + J_{AE}m_E(t))\right)$$

$$m_E(t + 1) = \tanh(J_{EA}m_A(t)) \tag{12}$$

For simplicity, we study the case where agent–environment connections are symmetric $J_{AE} = J_{EA} = J_c$, and $J_{AA} = J_r$. We numerically compute that the system has a similar solution than in the previous case, presenting a pitchfork bifurcation at a critical point (Fig. 3.B,D).

Moreover, we compute the value of $\varphi_{\mathcal{M}}(\tau \to \infty)$ for different mechanisms. For the case of the mechanism covering the whole system $\mathcal{M} = AE$, we look for the MIP of the system by isolating single units of the mechanism at $s(t)$ (Fig. 1.B). If we isolate a unit from region A, two connections are cut (one with value J_r and one with value J_c). Otherwise, if we isolate a unit from region E, only one connection with value J_c is cut. Thus, this second partition is always the MIP of the system (MIP_{AE}). For $\mathcal{M} = A$, the only candidate for the MIP is isolating one node from A, therefore cutting one connection with value J_r (MIP_A). Finally, for mechanism E there are no connections within the mechanism and we can directly conclude that $\varphi_E = 0$.

Now, the question is: can we consider A as an individual system or should we consider instead the coupled system AE as an integrated unit? Assuming $r_A = r_E = 0.5$, we define the values of integrated information as:

$$\varphi_A = \frac{1}{4}(|\sum_{\mathcal{R} = A,E} F'_{\underset{MIP_A}{\mathcal{R}}}(m_0, \tau, 0)|)$$

$$\varphi_{AE} = \frac{1}{4}(|\sum_{\mathcal{R} = A,E} F'_{\underset{MIP_{AE}}{\mathcal{R}}}(m_0, \tau, 0)|) \tag{13}$$

In Fig. 3.C,E we estimate the value of φ_A, φ_{AE} for $\tau \to \infty$ an initial value m_0 corresponding to the stationary solution of the system, and values of $J_c = 0.8$ (left) and $J_c = 1.2$ (right). We observe that in all cases the values of φ_A, φ_{AE} diverge next to the critical point. Nevertheless, in the first case when agent–environment connections are weaker $\varphi_A > \varphi_{AE}$ next to the critical point. In contrast, for stronger couplings between agent and environment $\varphi_A < \varphi_{AE}$ in the vicinity of the critical point.

We validate this result by solving Eq. (12) near criticality. We do this by transforming it into a system of one equation $m_A = \tanh(\frac{1}{2}(J_{AA}m_A + J_{AE}\tanh(J_{EA}m_A)))$ and finding its Taylor series near $m_A = 0$. We obtain that near the critical point:

$$F_A(m_0, \tau \to \infty, 0) = \sqrt{\frac{3(J_{AA} + J_{AE}J_{EA} - 2)}{J_{AE}J_{EA}^3 + \frac{1}{4}(J_{AA} + J_{AE}J_{EA})^3}}$$

$$F_E(m_0, \tau \to \infty, 0) = \tanh(J_{EA}F_A(m_0, \tau \to \infty, 0)) \tag{14}$$

Similarly, $F_A(m_0, \tau \to \infty, x)$ and $F_E(m_0, \tau \to \infty, x)$ are easily calculated by adding a $(1 - x)$ factor to the partitioned connections. Thus, we find that the location of the critical point which is the one satisfying $J_{AA} + J_{AE}J_{EA} = 2$ (Fig. 3.F). From here, we get:

$$F'_{A|MIP_A} = \frac{3}{2}\frac{-J_{AA}}{J_{AE}J_{EA}^3 + \frac{1}{4}(J_{AA} + J_{AE}J_{EA})^3}(\frac{1}{4}(J_{AA} + J_{AE}J_{EA})^2 F_A + \frac{1}{F_A})$$

$$F'_{E|MIP_A} = \frac{J_{EA}}{\cosh(J_{EA}F_A)^2}F'_{A|MIP_A}$$

$$F'_{A|MIP_{AE}} = \frac{3}{2}\frac{-J_{AE}J_{EA}}{J_{AE}J_{EA}^3 + \frac{1}{4}(J_{AA} + J_{AE}J_{EA})^3}(\frac{J_{EA}^2}{3} + \frac{1}{4}(J_{AA} + J_{AE}J_{EA})^2 F_A + \frac{1}{F_A})$$

$$F'_{E|MIP_{AE}} = \frac{J_{EA}}{\cosh(J_{EA}F_A)^2}F'_{A|MIP_{AE}}$$

where $F_{\mathcal{R}} = F_{\mathcal{R}}(m_0, \tau \to \infty, 0)$ and $F'_{\mathcal{R}|MIP_S} = F'_{\underset{MIP_S}{\mathcal{R}}}(m_0, \tau \to \infty, 0)$.
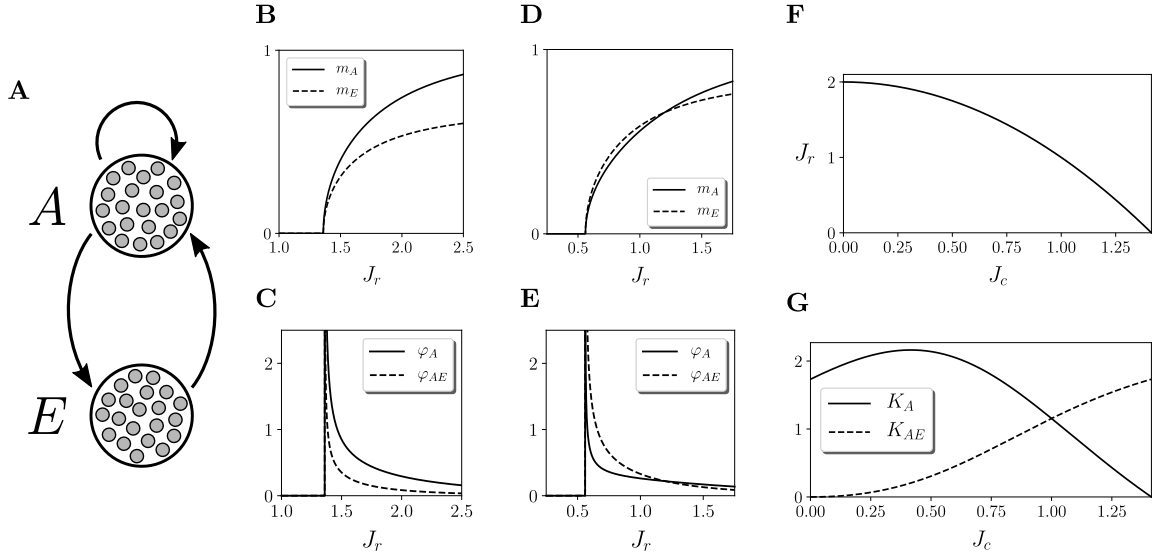
Fig. 3. Asymmetric interaction in a kinetic Ising model. A: Basic agent connected to an environment. B, C, D, E: Values of the mean fields (top, only positive values are shown) of the stable solution and integrated information $\varphi(\tau \to \infty)$ (bottom) for the agent and environment nodes of the model at stability for $J_c = 0.8$ (left) and $J_c = 1.2$ (right) and different values of J_r. F: Position of the critical point in the parameter space for different combinations of J_r, J_c. G: Constants multiplying $\varphi_A(\tau \to \infty)$ and $\varphi_{AE}(\tau \to \infty)$ near the critical point, showing the level of irreducibility of the parts of the system.

Near the critical point at $(J_{AA} + J_{AE}J_{EA}) \to 2^+$, the values of integrated information are approximated by the expressions:

$$
\begin{aligned}
\varphi_A &= J_{AA}K(J_{AA} + J_{AE}J_{EA} - 2)^{-1/2}, \\
\varphi_{AE} &= J_{AE}J_{EA}K(J_{AA} + J_{AE}J_{EA} - 2)^{-1/2}, \\
K &= \frac{\sqrt{3}(1 + J_{EA})}{\sqrt{J_{AE}J_{EA}^3 + \frac{1}{4}(J_{AA} + J_{AE}J_{EA})^3}}
\end{aligned}
\tag{15}
$$

by defining $K_A = J_{AA}K$ and $K_{AE} = J_{AE}J_{EA}K$ we describe with these variables the level of integrated information of the agent and the whole agent–environment system near the critical point. In Fig. 3.G we observe that there is a transition from the agent being the system with highest integration to the agent–environment. This illustrates that, near a critical point, the value of integrated information scales up indefinitely in an agent–environment system. In the case of symmetric interaction only for some cases the agent can be identified as the predominant integrated unit in the system, while in others the agent–environment system is the predominant unit.

3.3. Adaptive integrated information in the face of environmental diversity

We have just used integrated information for exploring the conditions that delimit an agent interacting with a static environment. The environment was 'passive' in the sense that it showed no self-interaction. This is not a common scenario, since typically environments change and display their own dynamics. A key aspect of agency is the ability of an agent to sometimes *modulate* the coupling with its environment to preserve its integrity (Barandiaran et al., 2009), generating an *interactional asymmetry* between agent and environment. Thus, a basic feature of living and cognitive systems is to display adaptive mechanisms regulating its coupling to the environment to maintain their level of functional integration for a range of external environments.

In order to characterize a scenario that is more realistic in this sense, we model an agent with two internal regions A and B, interacting with an environment E with recurrent connections (Fig. 4.A). A and B present feedback loops that we fit in order to maintain integration for a range of environmental parametric configurations. The evolution of the system is described by:

$$
\begin{aligned}
m_A(t+1) &= \tanh(\tfrac{1}{3}(J_{AA}m_A(t) + J_{AB}m_B(t) + J_{AE}m_E(t))) \\
m_B(t+1) &= \tanh(\tfrac{1}{3}(J_{BA}m_A(t) + J_{BB}m_B(t) + J_{BE}m_E(t))) \\
m_E(t+1) &= \tanh(\tfrac{1}{3}(J_{EA}m_A(t) + J_{EB}m_B(t) + J_{EE}m_E(t)))
\end{aligned}
\tag{16}
$$

We assume that the environment is defined by two parameters defining the agent–environment coupling $J_{AE} = J_{BE} = J_{EA} = J_{EB} = J_c$ and environmental self-coupling $J_{EE} = 1$. Values of $J_{AA}, J_{AB}, J_{BA}, J_{BB}$ will be tuned to maximize integration. We also assume $r_S = r_M = r_E = 1/3$.

In particular, the system will be tuned to maximize the integrated information of the agent AB, φ_{AB} while facing 5 different environments defined by values of J_c uniformly distributed in the range [0.8, 1.2]. We calculate φ for different parameters as in previous cases, testing the possible candidates for the MIP (in the case of φ_{AB}, the MIP candidates are isolating one node either from A or B) and the one minimizing integrated information is chosen.

In order to find the parameter values that maximize φ_{AB} for the set of environments, we first run a microbial genetic algorithm (Harvey, 2009) and then (using the parameters of the agent with larger fit) a Nelder–Mead algorithm (Nelder & Mead, 1965) to adjust the results. For both algorithms, the fitness function is defined as the value of $\varphi_{AB}(\tau)$, with some exceptions. For reducing the computational cost, the value of τ will be 10^4 for the genetic algorithm and 10^5 for the Nelder–Mead algorithm. In order to avoid the case where A and B are independent integrated units, fitness will be set to zero in the case that φ_A or φ_B are larger than φ_{AB}. As well, fitness is set to zero in the case where φ_{AB} does not converge to a stationary value.

After running the genetic and Nelder–Mead algorithms, we obtain an agent with parameters $J_{AA} = 0.09973671$, $J_{AB} = -0.85774749$, $J_{BA} = -0.8995672$ and $J_{BB} = 0.14326043$. This agent presents negative weights connecting A and B and positive self-coupling values. Thus, each region will inhibit the behaviour of the other while reinforcing itself, therefore regulating its activity to maintain high integrated information for the presented environments.
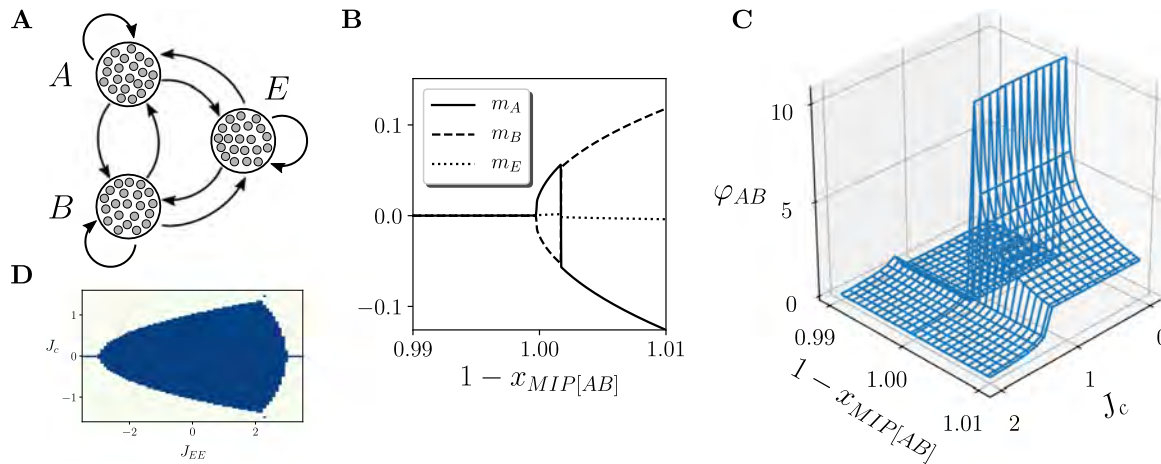
Fig. 4. Adaptive integration in a kinetic Ising model. A: Adaptive sensorimotor system connected to an environment. B: Values of the mean fields of the stable solution for a $J_c = 1$. C: Values of $\varphi_{AB}(\tau \to \infty)$ for different values of J_c. F: The blue area represents the surface in J_c and J_{EE} where $\varphi_{AB}(\tau \to \infty)$ diverges.

After tuning the parameters of the system, we evaluate its behaviour for different environments. For the values of J_c used during training, we find that the mean values of regions A and B, m_A and m_B display a similar transition than the previous examples (Fig. 4.B shows the case of $J_c = 1$, although other cases are similar). Moreover, we can observe that there is a divergence of the values of φ_{AB} for a range of values of J_c (Fig. 4.C). For larger values of J_c the transition disappears and the values of φ_{AB} do not diverge.

The example presented here displays an important qualitative change in comparison with the previous one. The value of φ_{AB} diverges but not only for a specific environment due to fine tuning of its self-couplings as in the previous case. Instead, the divergence is maintained for an approximate range of values of J_c of $[-1.21, 1.21]$. Moreover, this divergence is also maintained if we modify the value of J_{EE}, displaying a surface in which the value of $\varphi(\tau)$ diverges (Fig. 4.D). This means that the points of divergence from previous examples are transformed here into a critical surface that maintains integration of the system for a wide range of environmental parameters. That is, the system is able to self-regulate to some extent to maintain its integration, and thus its viability as an agent.

4. Discussion

We have proposed a modified measure of integrated information φ which, together with mean field approximations in a kinetic Ising model, allows us to capture for the first time integrated information in very large neural networks, up to the thermodynamic limit. Using this method we are able to compute φ for infinite size mean field kinetic Ising models with quasi-homogeneous infinite-range connectivity.

Our models, although highly idealized, allow us to speculate about some of the properties of integrated neural organization. First, we observe that, despite the infinite size of the models, the amount of integrated information is bounded for most of its parameter space. Only at critical points does the level of total integrated information diverge, suggesting that integrated networks need to be poised near critical points in their parameter space to maintain their level of integration as their size grows. This suggests that it may be of greater interest to describe brain organization in terms of diverging tendencies of integration in different modules rather than in terms of the specific values of φ in finite systems.

Furthermore, we have shown that integrated information can be used to define the boundaries between a system and its

environment by comparing the divergent tendencies of their joint and individual integration. For doing so, some of the assumptions of current formulations of IIT had to be modified. An important implication of our model is that integrated information cannot, in principle, be measured in a brain independently of its environment (bodily and extra-bodily), nor by assuming that the environment is an independent source of noise. Moreover, our results show that near critical points in some cases both the system and system–environment integrated information diverges. Nevertheless, we have shown a way to characterize the dominant dynamical unit by comparing the difference in the diverging tendencies between the two configurations.

Our results connect the emergence of boundaries of integration with phenomena related to criticality. Systems near critical points are maximally sensitive to changes in some directions of their parameter space (sensitivity is generally measured as the susceptibility of the system to changes in a given parametrical direction). Here, we capture integrated information measures by applying different partitions to the system which are interpreted as changes in particular directions of the parameter space. Thus, the level of integrated information corresponds to the susceptibility of the system to the minimum information partition, i.e., the partition with the less significant effect on the system's causal powers. In the framework of IIT, systems highly sensitive to their minimum information partition are interpreted as maximally irreducible units.

This could allow further simplifications in order to measure integrated information in complex models or even empirical setups. By testing the behaviour of a system when perturbations in its components are introduced (i.e., noise injected in partitioned connections), the integrated information of a mechanism can be described as the minimal susceptibility to perturbations from the set of possible partitions. The connection between information integration and critical susceptibility allows us to speculate about the link between integration and properties that have been postulated as pervasive of living beings such as self-organized criticality (Bak, Tang, & Wiesenfeld, 1988).

By interpreting integrated information in terms of susceptibilities in the parametrical direction of partitions of the system, we can think of integration as the sensitivity of a system to the decoupling of the modules composing it. In our last example, we show that internal regulation results in the capacity for maintaining this susceptibility for a range of different situations. We hypothesize that this can be achieved by similar dynamics as those of systems showing self-organized criticality,

which are attracted to critical points of maximum susceptibility. Such systems would be capable of self-organizing near points where they can maintain maximal sensitivity to the integrity of their internal organization while interacting with changing environments (e.g., maintaining internal invariances near critical surfaces (Aguilera & Bedia, 2018)).

5. Conclusion

The core ideas that IIT intends to capture apply to a variety of poorly understood questions in biological, neural, and cognitive processes. By introducing some modifications to the assumptions of the theory to take into account different temporal spans and influences from the environment, and studying the behaviour of integration measures in the thermodynamic limit, we have shown the existence of critical points that maximize a system's integration, for instance, an organism or a cognitive agent. The fact that our case studies remain general and abstract (we do not specify any detail about the neural, sensorimotor, and environmental processes involved) suggests that robust individuation and susceptibility towards loss of integration are inherent consequences of maximizing a tendency towards integration, and so they are likely to be observable trends in all systems that are able to do so.

A limiting assumption in our approach is the homogeneity of the elements within a each region. Biological networks cannot be assumed to present such a degree of homogeneity and the variability in their components and interactions has to be accounted for. Our framework, however, can take into account higher levels of heterogeneity by introducing a larger number of regions. In the case of three regions we observe that tuning the parameters of the system results in the extensions of critical points of diverging integration into regions of the parameter space. We expect (but have not yet verified) that increasing the number of interacting regions will still result in critical regions of divergent integration. In neural network models, it has been found that structural heterogeneity can generate extended critical-like regions (Moretti & Muñoz, 2013), thus we may also expect this phenomenon to be reinforced in the presence of higher heterogeneity in our models. Our results are also limited to models with stationary solutions where we can evaluate the stable solution when the temporal span tends to infinity. This is not a limitation of the method, though. The results of more realistic systems presenting cyclic or chaotic dynamics could be harder to interpret, although they are in principle tractable within the framework presented here and could be explored in further work.

Despite these simplifications, our models move beyond some idealized assumptions of the original IIT framework and lead to considerations that are hard to formulate otherwise. In this sense, the models allow a shift of focus toward the integrative tendencies of systems as they grow or evolve. This different set of idealizations opens up the applicability of IIT to a range of questions about changes over developmental and evolutionary time. Even in the simple cases we have considered, the existence of critical points that maximize integration may be important for understanding apparent jumps in complexity, including the transitions at the origin of life (Walker & Davies, 2013) or cognitive developmental transitions (Molenaar & van der Maas, 2004).

Focusing on the divergent tendencies of integration measures, we are able to capture the asymmetry of agent–environment interactions. Thinking about interactions with the environment in this terms is fruitful for grounding notions such as the individuality or the autonomy of a system. Often, these concepts have been formalized in terms of self-determination and independence from an environment (Bertschinger, Olbrich, Ay, & Jost, 2008; Krakauer, Bertschinger, Olbrich, Ay, & Flack, 2014). By contrast,

our examples show how both integration of a system and integration between system an environment can diverge together, while the level of individuality of the system can be quantified by their relative divergence speed. This is a robust finding obtained under the minimal assumptions and thus, we suggest, a general trend in large complex systems. The key data of interest as systems scale up are not so much the absolute values of integrated information, but the relative divergent tendencies of system integration and system–environment integration.

In addition, by exploring different kinds of agent–environment configurations, we observe that agents assumed to maximize integration are likely to do so robustly for a range of environmental situations due to the existence of critical surfaces. The existence of these surfaces that guarantee maximal integration is coherent with postulates in the theoretical foundations of adaptive systems research, such as the existence of 'regions of viability' that guarantee the integrity of an agent (Ashby, 1960; Barandiaran & Egbert, 2014). While such conditions of viability have often been imposed by the designer or assumed to be given by evolutionary or material constraints, our approach allows us to think of them as critical regions emerging at the level of the integrative forces of the system. This illustrates how viability regions could scale up from material or pre-given constraints to regions defined by increasing complexity of the integrated activity of a system.

Acknowledgements

Appendix A. IIT 3.0

In the last version of integrated information theory (Oizumi et al., 2014), integrated information of a subset of elements of a system is computed as follows. For a system of elements S in state s, we describe the input–output relationship of the system elements through its corresponding transition probability function p, describing the probabilities of the transitions from one state to another for all possible system states. IIT requires that p satisfies the Markov property (i.e., the state at time t only depends on the state at time $t-1$), and that the current states of elements are independent, conditional on the past state of the system. This conditions are satisfied by the asymmetric kinetic Ising model used here.

For any two subsets of S, called the mechanism $\mathcal{M}$ and the purview $\mathcal{P}$, we can define the cause and effect repertoires of $\mathcal{P}$ over $\mathcal{M}$, that is, how $\mathcal{M}$ in its current state $\{s_i(t)\}_{i \in \mathcal{M}}$, constrains the potential past or future states of $\{s_i(t-1)\}_{i \in \mathcal{P}}$ or $\{s_i(t+1)\}_{i \in \mathcal{P}}$. Cause and effect repertoires of the system are described by the probability distributions $p(\mathcal{P}_{t-1}|\mathcal{M}_t) = p(\{s_i(t-1)\}_{i \in \mathcal{P}}|\{s_i(t)\}_{i \in \mathcal{M}})$ and $p(\mathcal{P}_{t+1}|\mathcal{M}_t) = p(\{s_i(t+1)\}_{i \in \mathcal{P}}|\{s_i(t)\}_{i \in \mathcal{M}})$.

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

$$\varphi_{cause} = \max_{\mathcal{P}} \left(\min_{cut} \left(D(p(\mathcal{P}_{t-1}|\mathcal{M}_t), p^{cut}(\mathcal{P}_{t-1}|\mathcal{M}_t)) \right) \right)$$

$$\varphi_{effect} = \max_{\mathcal{P}} \left(\min_{cut} \left(D(p(\mathcal{P}_{t+1}|\mathcal{M}_t), p^{cut}(\mathcal{P}_{t+1}|\mathcal{M}_t)) \right) \right)$$
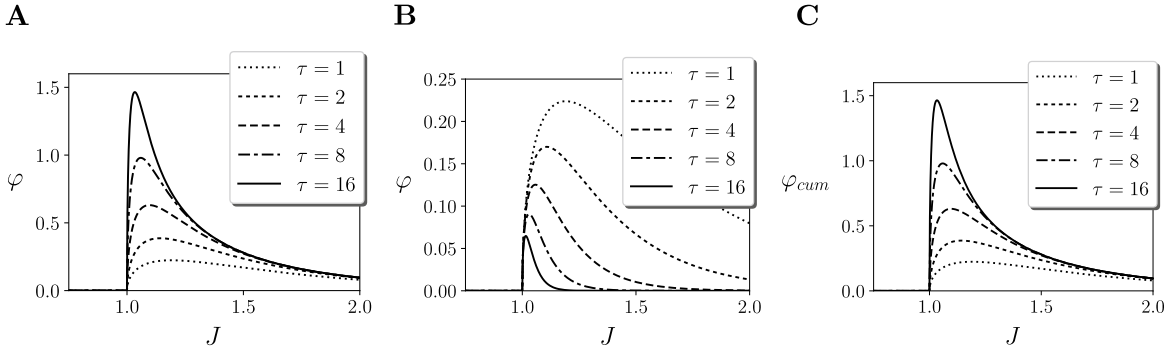
(A.1)

A **B** **C**

Fig. B.1. Temporal range of integration. A: Values of $\varphi(\tau)$ using continuous injection of noise for different values of J. B: Values of $\varphi(\tau)$ using an initial injection of noise for different values of J. C: Values of $\varphi_{cum} = \sum_{\tau'=1}^{\tau} \varphi(\tau')$ using an initial injection of noise for different values of J.
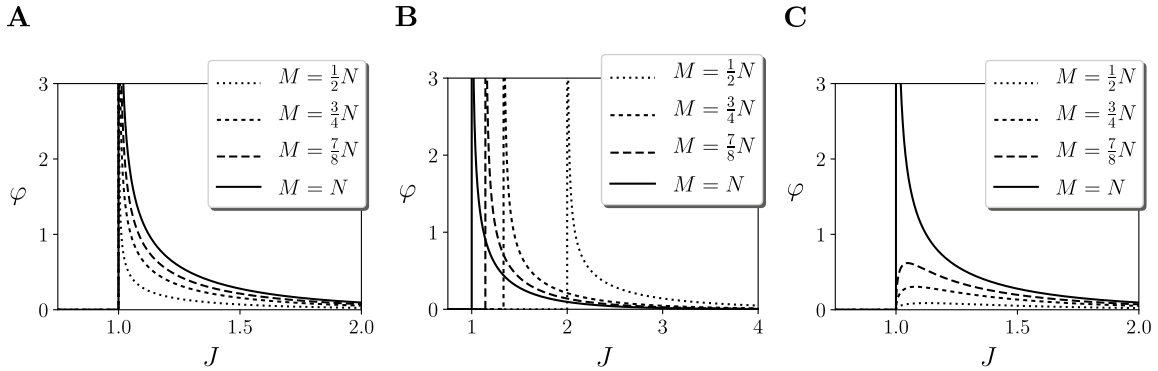
A **B** **C**

Fig. B.2. Effects of the environment in integrated information. Values of $\varphi_{\mathcal{M}}(\tau \to \infty)$ of a mechanism $\mathcal{M}$ of size M for different values of J, assuming that elements outside of the mechanism operate (A) normally, (B) as independent sources of noise, and (C) as static input fields.

where *cut* is a partition of the mechanism into two halves, and p^{cut} the cause or effect probability distribution under the partition,

$$cut = \{\mathcal{M}_1, \mathcal{P}_1, \mathcal{M}_2, \mathcal{P}_2\}$$
$$p^{cut}(\mathcal{P}|\mathcal{M}) = p(\mathcal{P}_1|\mathcal{M}_1) \otimes p(\mathcal{P}_2|\mathcal{M}_2) \tag{A.2}$$

The integrated information of the mechanism $\mathcal{M}$ is the minimum of its corresponding integrated cause and effect information,

$$\varphi = min(\varphi_{cause}, \varphi_{effect}) \tag{A.3}$$

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

$$\Phi = \min_{cut} D(C, C^{cut}) \tag{A.4}$$

For both the integrated information of a mechanism (φ) and the integrated information of a system (Φ), distance D is computed as the Wasserstein or earth mover's distance. Finally, if S is a subset of elements of a larger system, all elements outside of S are considered as part of the environment and are conditioned on their current state throughout the causal analysis. Further details of the steps described here can be found in Oizumi et al. (2014).

Appendix B. Simplified integrated information φ

Measures in this paper are inspired by the IIT framework, although we apply some modifications and simplifications. In this appendix we detail the changes respect our version of φ and the one described by Oizumi et al. (2014).

B.1. Temporal range

First, as we mentioned in the paper, we only compute the value of φ for the effects of the current system in a posterior state $t + \tau$, while IIT computes the minimum of φ_{cause} and φ_{effect} at $t - 1$ and $t + 1$. However, IIT can also deal with temporal scales. As IIT operates with the transition probability matrix of a system, one could compute this matrix from time t to time $t + \tau$ and apply the operations for computing φ over it. This implies that the noise introduced by partitions in the connections that are cut down is only injected at time t, and the system behaves normally for the following steps. In our case, we inject independent noise at every update from time t to $t + \tau$.

We can test the difference between the two approaches in a homogeneous kinetic Ising model with $H_i = 0$ and $J_{ij} = J$. As we showed in the paper, applying a continuous noise injection in partitions makes the value of φ diverge around the critical point $J = 1$ as τ grows (Fig. B.1.A). Conversely, in we only apply an initial noise injection at partitioned connections, we see that the measured φ operates in a different way (Fig. B.1.B). In this case, as τ increases, the value of φ decreases as the system regains stability in its original position. Moreover, for small values of τ the values of J with larger φ are above the critical point. However, we observe that, the closer we are to the critical point, the slower φ decreases. This is due to a phenomena called 'critical slowing down', a phenomena characteristic of critical dynamics in which the response time of a system near criticality tends to infinity. Curiously, if we compute the cumulative sum of the values of φ from 1 to τ, i.e. $\varphi_{cum} = \sum_{\tau'=1}^{\tau} \varphi(\tau')$ (Fig. B.1.C), we observe that the result is identical to the case of continuous noise injection at partitions.

B.2. Purview

In IIT, integrated information of a mechanism $\varphi_{\mathcal{M}}^{MIP}$ is evaluated not only for a particular mechanism $\mathcal{M}$, but also for a purview $\mathcal{P}$. If the mechanism defines which units of $\{s_i(t)\}_{i \in \mathcal{M}}$ we take into account, the purview defines which units of the future state $\{s_i(t+\tau)\}_{i \in \mathcal{P}}$ we take into account. Given these subset of present and future states, partitions are computed over the join space of $\{s_i(t)\}_{i \in \mathcal{M}}$ and $\{s_i(t+\tau)\}_{i \in \mathcal{P}}$, and the purview $\mathcal{P}$ with maximum integrated information for its MIP is selected. Here for simplicity, we apply the partition over $\{s_i(t)\}_{i \in \mathcal{M}}$ and $\{s_i(t+\tau)\}_{i \in \mathcal{M}}$, making the mechanism and purview coincide, and 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.

Allowing more choices of purview could make a big difference in certain systems, although in the quasi-homogeneous systems tested in the paper the differences are small.

B.3. Elements outside of a mechanism

More importantly, there are significant differences from the IIT framework in the way we treat the elements that are outside of the evaluated mechanism $\mathcal{M}$. In IIT, elements outside the mechanism are assumed to be unconstrained (i.e., as random as possible). We decided to modify this assumption because it can have dramatic effects when measuring the behaviour of large systems. Specifically, assuming unconstrained elements outside the mechanism create an artifact that provokes a shift in the critical point of the system (this will be detailed in future work).

Let us provide an example using a homogeneous Ising model with local fields $H_i = 0$ and couplings $J_{ij} = J$. As we shown, compute the value of φ for the whole system using continuous noise injection at partitioned connection yields a divergence around the critical point at $J = 1$. Now, we will show what is the behaviour of its internal mechanisms assuming different behaviours of the units outside of the mechanism.

First, we compute values of mechanism covering a fraction of the system M/N (since the system is homogeneous, any fraction we choose has the same behaviour) assuming that the elements outside of the mechanism $\mathcal{M}$ keep operating normally (Fig. B.2.A). In this case, we observe that the divergence of $\varphi_{\mathcal{M}}$ is maintained, although the value of $\varphi_{\mathcal{M}}$ decreases with the mechanism size.

In contrast, if we accept IIT assumption and take the elements of the mechanism as independent sources of noise, the behaviour of $\varphi_{\mathcal{M}}$ changes radically. In this case, the divergence is maintained but takes place at a different value of the parameter J (Fig. B.2.B). This happens because independent sources of noise have a zero mean field value, and thus the phase transition of the system takes place at larger values of J that compensate the units that now are contributing with a zero mean field. Thus, considering the elements outside of the mechanism as independent sources of noise can be misleading about the operation of mechanisms that are embedded in large systems.

A less loaded assumption could be maintaining the state of the units outside of the mechanism with the static values that they had at time t, that is, maintaining their mean field constant. We can see at Fig. B.2.C that this behaviour is also not satisfactory, since for mechanism sizes smaller than N the value of $\varphi_{\mathcal{M}}$ decreases very rapidly, and it is exactly zero at the critical point. We can understand this thinking that the effect of constant fields is equal to adding a value of H_i equal to the input from frozen units, therefore breaking the symmetry of the system and precluding a phase transition.

B.4. Mean field approximation of partitioned systems

We simplify the calculation of the probabilities $p(\{s_i(t+\tau)\}_{i \in \mathcal{M}} | \{s_i(t)\}_{i \in \mathcal{M}})$ and $p^{cut}(\{s_i(t+\tau)\}_{i \in \mathcal{M}} | \{s_i(t)\}_{i \in \mathcal{M}})$ by using a mean field approximation described by Eqs. (3) and (4).

In the case of partitioned systems for computing integrated information, cutting connections injects uniform noise on the input node. In the mean field approximation, this would be equivalent to inject a zero mean field signal, which is equivalent to setting to zero the affected connection weights when computing $h_i(t)$.

B.5. Integrated conceptual information

Finally, once φ is computed, IIT proposes a second level of calculations for computing integrated conceptual information Φ where new bidirectional partitions are applied to the system. In our case, given the homogeneity of the system, we do not compute conceptual information since all the mechanisms composing each set have similar behaviour. Thus, for simplicity we do not apply a second level of partitions.

References

Aguilera, M., & Bedia, M. G. (2018). Adaptation to criticality through organizational invariance in embodied agents. *Scientific Reports*, 8(7723).

Aguilera, M., Bedia, M. G., Santos, B. A., & Barandiaran, X. E. (2013). The situated HKB model: how sensorimotor spatial coupling can alter oscillatory brain dynamics. *Frontiers in Computational Neuroscience*, 7(117).

Albantakis, L., Hintze, A., Koch, C., Adami, C., & Tononi, G. (2014). Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Computational Biology*, 10(12), e1003966.

Ashby, W. R. (1960). *Design for a brain: The origin of adaptive behavior.* New York: Wiley.

Bak, P., Tang, C., & Wiesenfeld, K. (1988). Self-organized criticality. *Physical review A*, 38(1), 364.

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5), 367–386.

Barandiaran, X. E., & Egbert, M. D. (2014). Norm-establishing and norm-following in autonomous agency. *Artificial Life*, 20(1), 5–28.

Barrett, A. B., & Seth, A. K. (2011). Practical measures of integrated information for time-series data. *PLoS Computational Biology*, 7(1), e1001052.

Bassett, D. S., & Gazzaniga, M. S. (2011). Understanding complexity in the human brain. *Trends in Cognitive Sciences*, 15(5), 200–209.

Bertschinger, N., Olbrich, E., Ay, N., & Jost, J. (2008). Autonomy: An information theoretic perspective. *Biosystems*, 91(2), 331–345.

Harvey, I. (2009). The microbial genetic algorithm. In *Lecture notes in computer science, Advances in artificial life. Darwin meets Von Neumann* (pp. 126–133). Springer, Berlin, Heidelberg.

Hoel, E. P., Albantakis, L., Marshall, W., & Tononi, G. (2016). Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, 2016(1).

Kitazono, J., Kanai, R., & Oizumi, M. (2018). Efficient algorithms for searching the minimum information partition in integrated information theory. *Entropy*, 20(3), 173.

Krakauer, D., Bertschinger, N., Olbrich, E., Ay, N., & Flack, J. C. (2014). The information theory of individuality. arXiv preprint arXiv:1412.2447.

Marshall, W., Albantakis, L., & Tononi, G. (2018). Black-boxing and cause-effect power. *PLoS Computational Biology*, 14(4), e1006114.

Marshall, W., Kim, H., Walker, S. I., Tononi, G., & Albantakis, L. (2017). How causal analysis can reveal autonomy in models of biological systems. *Philosophical Transactions of Royal Society A*, 375(2109), 20160358.

Mediano, P. A. M., Seth, A. K., & Barrett, A. B. (2019). Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. *Entropy*, 21(17).

Molenaar, P. C. M., & van der Maas, H. L. J. (2004). Commentary on: "Piaget's stages: The unfinished symphony of cognitive development" by D.H. Feldman. *New Ideas in Psychology*, 22.

Moretti, P., & Muñoz, M. A. (2013). Griffiths phases and the stretching of criticality in brain networks. *Nature Communications, 4*, 2521.

Mudrik, L., Faivre, N., & Koch, C. (2014). Information integration without awareness. *Trends in Cognitive Sciences*, 18(9), 488–496.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology*, *10*(5), e1003588.

Oizumi, M., Amari, S. -i., Yanagawa, T., Fujii, N., & Tsuchiya, N. (2016). Measuring integrated information from the decoding perspective. *PLoS Computational Biology*, *12*(1), e1004654.

Oizumi, M., Tsuchiya, N., & Amari, S. -i. (2016). Unified framework for information integration based on information geometry. *Proceedings of the National Academy of Sciences*, *113*(51), 14817–14822.

Pessoa, L. (2014). Understanding brain networks and brain organization. *Physics of Life Reviews*, *11*(3), 400–435.

Pressé, S., Ghosh, K., Lee, J., & Dill, K. A. (2013). Principles of maximum entropy and maximum caliber in statistical physics. *Reviews of Modern Physics*, *85*(3), 1115–1141.

Schneidman, E., Berry, M. J., Segev, R., & Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, *440*(7087), 1007–1012.

Tononi, G., & Edelman, G. M. (1998). Consciousness and complexity. *Science*, *282*(5395), 1846–1851.

Varela, F. J. (1995). Resonant cell assemblies: A new approach to cognitive functions and neuronal synchrony. *Biological Research*, *28*(1), 81–95.

Walker, S. I., & Davies, P. C. W. (2013). The algorithmic origins of life. *Journal of the Royal Society Interface*, *10*(79).