Integrated Information and Autonomy in the Thermodynamic Limit

Miguel Aguilera[1,2], Ezequiel A. Di Paolo[1,3]

[1]IAS-Research Center for Life, Mind, and Society, University of the Basque Country, Donostia, Spain
[2]ISAAC Lab, Aragón Institute of Engineering Research, University of Zaragoza, Zaragoza, Spain
[3]Ikerbasque, Basque Foundation for Science, Bizkaia, Spain
sci@maguilera.net

Abstract

The concept of autonomy is fundamental for understanding biological organization and the evolutionary transitions of living systems. Understanding how a system constitutes itself as an individual, cohesive, self-organized entity is a fundamental challenge for the understanding of life. However, it is generally a difficult task to determine whether the system or its environment has generated the correlations that allow an observer to trace the boundary of a living system as a coherent unit. Inspired by the framework of integrated information theory, we propose a measure of the level of integration of a system as the response of a system to partitions that introduce perturbations in the interaction between subsystems, without assuming the existence of a stationary distribution. With the goal of characterizing transitions in integrated information in the thermodynamic limit, we apply this measure to kinetic Ising models of infinite size using mean field techniques. Our findings suggest that, in order to preserve the integration of causal influences of a system as it grows in size, a living entity must be poised near critical points maximizing its sensitivity to perturbations in the interaction between subsystems. Moreover, we observe how such a measure is able to delimit an agent and its environment, being able to characterize simple instances of agent-environment asymmetries in which the agent has the ability to modulate its coupling with the environment.

Introduction

Many open questions in biology are related to major evolutionary transitions in biological organization. How do living systems arise from non-living matter? How does biological individuality emerge from networks of complex chemical reactions? How do autonomous agents constitute themselves in front of their environment? Understanding the difference between the living and the non-living, the cognitive and non-cognitive, implies understanding how these transitions work. It has been proposed that the transitions related with the origin of life may be akin to physical transitions (as thermodynamic phase transitions) associated with a shift in the causal structure of a system (Walker and Davies, 2013)

The challenge is to define and quantify the type of organization that emerges in these transitions. Although the number of open questions regarding this issue is dauntingly large, one property of living organization that has sparked interest during the last decades is the idea of autonomy. Roughly, autonomous systems can be described as forming a unitary whole that emerges from the interaction of its components in a self-organized manner (Maturana and Varela, 1980; Moreno and Mossio, 2015). That is, they are able to preserve themselves as entities with self-defined and self-maintained boundaries while facing internal and external perturbations.

Although the idea of autonomy presents an exciting theoretical perspective, in practice quantifying autonomy presents some formidable difficulties. Many of these difficulties are related to the fact that autonomy requires a system to constitute itself as a unified whole that can be regarded as distinct from the environment, yet in ongoing interaction with it. Hence, the question lies in how to distinguish between a living entity and its environment that are constituted as distinct entities at a macroscopic level, while remaining meshed together in the material interactions at the microscopic level.

Recent efforts to characterize autonomous organization have used information theory in dynamical systems (Bertschinger et al., 2008). Nevertheless, these approaches, based on measuring correlations between variables at different times, present some limitations for distinguishing between system and environment. Typically, while these nonlinear correlations can be described in dynamical or information theoretical terms, it is not possible to capture what are the contributions of the system and environment, making the task to inferring the boundary between the two difficult.

Instead of analyzing mere correlations, it has been proposed that interventionist notions of causality are better suited to characterize autonomous organization (Marshall et al., 2017). That is, instead of assessing whether a system is unified into a coherent whole by analyzing its behaviour in stable conditions, one could capture the causal forces integrating the behaviour of the system by observing its behaviour under perturbations. Specifically, Marshall et al. (2017) have used the framework of *integrated information theory* (IIT, Oizumi et al., 2014) for delimiting the

boundaries of biological models. Although interventionist approaches have been used before in the analysis of complex systems, IIT is particularly interesting because it proposes a rigorous information theoretical framework for measuring the effects in interventions in the system in terms of the irreducibility of its components. Additionally, IIT does not require the existence of stationary conditions.

IIT postulates that any subset of elements of the system is a mechanism 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 the 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 in 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 bidiriectional 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 causal boundaries of the system are defined around the subset with larger Φ.

In general, IIT has been tested in small toy models (e.g. logic gates, Oizumi et al., 2014). Due to the computational complexity of IIT measures, it is not feasible to apply them to larger systems (some alternative formulations try to circumvent this problem, e.g. Oizumi et al., 2016). However, a serious attempt to capture aspects of biological autonomy should think about autonomy as an organizational property of a system, which is able to scale as it grows in size. In this article, we compute integrated information measures using mean field approximations in an infinite range Ising model. Our approach simplifies some aspects of the IIT framework and proposes some modifications in order to measure integrated information correctly as a system scales to very large sizes. We introduce a simple kinetic Ising model with quasi-homogeneous connectivity, which presents an exact mean field solution that we use to simplify the calculation of integrated information φ of the mechanisms of a system. As the model with infinite size has an exact mean field solution for some combinations of parameters, this allows us to test the application of ideas from IIT to systems near the thermodynamic limits. We argue that some minor changes are necessary in the current postulates of IIT to capture the diverging properties of very large systems. In addition, using some variations of the model we exemplify the case of determining the boundary of a system interacting with an external environment, which is described by comparing the difference between integrated information in the system alone and the joint agent-environment system. Our intent is not to propose a complete framework to measure integrated information, but to test how integrated information behaves with systems of diverging sizes, as well as to explore tentative routes to adapt IIT measures to these cases. We find that in some examples the specific value of integrated information of a system is not as informative as other properties as how it diverges with different spatial and temporal scales.

The results presented here represent a first attempt for using integrated information theory to delimit the causal boundaries of a family of infinite size systems that can be formally solved. The interest of the study is twofold. First, it allow us to check some of the assumptions of IIT and propose some modifications to maintain its consistency in the large thermodynamic limit, and to propose a way to adapt IIT measures in 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, as well as what type of transitions we might expect for understanding the evolutionary history of living organisms.

We proceed as follows. First, we introduce our model and a mean field approximation for solving it. Then we introduce integrated information theory and how it can be computed using mean field approximations, and illustrate this in a simple homogeneous model. Then, we present the results of our method in two scenarios in which the model is modified to represent asymmetric interactions of the type that we could expect from agent-environment interactions. Finally, we discuss the possible generalization and implications of our results.

Approximating Integrated Information for a Mean Field Kinetic Ising Model

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

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

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 can assume an $\beta = 1$.

Mean Field Approximation

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 $J_{ij} = \frac{1}{N_{\mathcal{R}}} J_{\mathcal{SR}}$, 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), a mean field approximation allows to calculate the field of all units i belonging to the region $\mathcal{S}$ as

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

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

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

Integrated Information

We use a simplified version of the integrated effect information described by IIT (Oizumi et al., 2014), implementing some modifications to correctly 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 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 $\{s_i(\tau_0)\}_{i \in \mathcal{M}}$. 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 (from the set of possible bipartitions) is applied over the units in $\mathcal{M}$. Figure 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 $s(\tau_0+\tau)$ systems is computed after τ updates injecting noise at the partitioned elements. Then, integrated information is defined as the distance D between the conditional probability distributions at $t + \tau$:

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

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 $\mathcal{M}$, $cut = \{S^c_1, S^c_2, S^f_1, S^f_2\}$, where S^c_1, S^c_2 design the blocks of a bipartition of the mechanism at the current state $\{s_i(t)\}_{i \in \mathcal{M}}$, and S^f_1, S^f_2 refer to the blocks of a bipartition (not necessarily the same) of the updated state of the units $\{s_i(t+1)\}_{i \in \mathcal{M}}$. Figure 1.B represents the partition $cut = \{\{s_1(t), s_2(t)\}, \{s_3(t)\}, \{s_1(t+1), s_2(t+1), s_3(t+1)\}, \{\}\}$.

Specifically, 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., $\varphi^{MIP}_M(\tau) = \min_{cut} \varphi^{cut}_M(\tau)$). We use $\varphi_{\mathcal{M}}(\tau)$ to denote the minimum information partition integrated information $\varphi^{MIP}_M(\tau)$.

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. To preserve the consistency of our results, we let elements outside the mechanism operate normally (see Appendix).

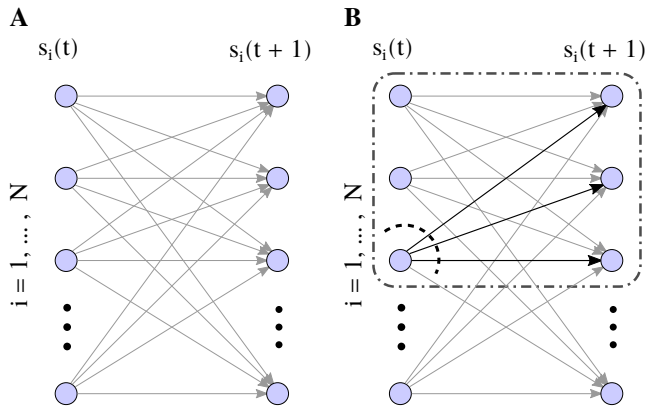

Figure 1: **(A)** Description of the infinite size kinetic Ising model. **(B)** Description of the partition schema used to define perturbations. Partitioned connections (black arrows) are injected with random noise.

Integrated Information in the Mean-Field Model

We now show how integrated information can be computed for the mean field approximation of 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 dis-

tribution dependent on the mean field at the previous step:

$$p(s(\tau_0+\tau)|s(\tau_0)) = \int p(s(\tau_0+\tau)|h(\tau_0+\tau))p(h(\tau_0+\tau)|s(\tau_0))dh(\tau_0+\tau) \quad (6)$$

Moreover, for the system of infinite size described above, the evolution of $h(t)$ is deterministic and governed by Equation 4, and given the mean field value received by each unit their posterior probability distribution is independent, so $p(s_i(\tau_0+\tau)|s(\tau_0)) = p(s_i(\tau_0+\tau)|h_i(\tau_0+\tau))$. In this context, the calculation of the Wasserstein distance D is drastically simplified, and we can compute φ as the sum of distances between independent binary variables:

$$\begin{aligned}\varphi^{cut}_{\mathcal{M}}(\tau) &= \sum_i D(p(s_i(\tau_0+\tau)|h_i(\tau_0+\tau)), \\ &\qquad p^{cut}(s_i(\tau_0+\tau)|h_i(\tau_0+\tau))) \\ &= \frac{1}{2}\sum_{\mathcal{R}} N_{\mathcal{R}}|m_{\mathcal{R}}(\tau_0+\tau) - m^{cut}_{\mathcal{R}}(\tau_0+\tau)|\end{aligned} \quad (7)$$

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 infinte 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 $m^{cut}_{\mathcal{R}}(\tau_0+\tau)$. Thus, finding the MIP corresponds to finding which region $\mathcal{R}$ of the system affects less future states when one node of the region is isolated in the partition at time t (e.g. Figure 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 Equation 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[1] 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:

$$\begin{aligned}&F^{\ }_{cut\ \mathcal{R}}(m_0, \tau, x) \\ &= F_{\mathcal{R}}(m_0, \tau, \{\{J_{\mathcal{S},\mathcal{R}}\}_{uncut}, \{(1-\frac{x}{r_{\mathcal{R}}})J_{\mathcal{S},\mathcal{R}}\}_{cut}\})|\end{aligned} \quad (8)$$

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 $F^{\ }_{cut\ \mathcal{R}}(m_0, \tau, 0)$ and $F^{\ }_{cut\ \mathcal{R}}(m_0, \tau, \frac{1}{N})$ respectively. Given this, assuming $N \to \infty$ we calculate the final form of φ as a sum of the derivatives of function $F^{\ }_{cut\ \mathcal{R}}(m_0, \tau, x)$:

$$\begin{aligned}&\varphi^{cut}_{\mathcal{M}}(\tau) \\ &= \lim_{N\to\infty}\frac{1}{2}\sum_{\mathcal{R}} N_{\mathcal{R}}|F^{\ }_{cut\ \mathcal{R}}(m_0, \tau, 0) - F^{\ }_{cut\ \mathcal{R}}(m_0, \tau, \frac{1}{N})| \\ &= \frac{1}{2}\sum_{\mathcal{R}}|r_{\mathcal{R}}F'^{\ }_{cut\ \mathcal{R}}(m_0, \tau, 0)|\end{aligned} \quad (9)$$

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.

Integrated Information in the 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 Figure 1.A). The system only has 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 Figure 2.B, the value of $\varphi_{\mathcal{M}_N}(\tau)$ appears to diverge as τ grows[2].

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 Figure 2.C we observe how $\varphi_{\mathcal{M}_N}(\tau \to \infty)$ 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 Figure 2.D, the resulting value of integrated information

[1] Note that cutting a connection implies injecting uniform noise, which in the mean field approximation is equivalent to substitute the input by a zero mean field or just removing the connection. This is an important approximation that allow 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.

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

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 $F_{cut}(m_0, \tau \to \infty, x)$ just by substituting $J \leftarrow J(1-x)$.

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 (Figure 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 the first two terms of its Taylor series, finding that in the limit $J \to 1^+$ we approximate:

$$\begin{aligned} F_{cut}(m_0, \tau \to \infty, x) &= \pm\sqrt{\frac{3(J(1-x)-1)}{(J(1-x))^3}} \\ \varphi_{\mathcal{M}_N}(\tau \to \infty) &= \frac{1}{2}\left|\frac{\sqrt{3}\,(2J-3)}{2\sqrt{J^3(J-1)}}\right| \end{aligned} \tag{10}$$

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. The value of integrated information per unit $\varphi_{\mathcal{M}_N}(\tau \to \infty)/N$ of the system would tend to 0 at any position but in the critical point. Thus, if a system must maintain its levels of integration per unit as it size increases, it may need to be poised near a critical point that shows a divergence of the values of φ.

Integrated Information for Measuring Agent-Environment Asymmetries

We apply the proposed measure of integrated information to the problem of determining the causal boundaries of an agent interacting with an environment. One of the central aspects of agency is the existence of agent-environment asymmetries (Barandiaran et al., 2009), in which the part of the system corresponding to the agent is able (to an extent) to define 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.

Bidirectional Agent-Environment 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 (Figure 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:

$$\begin{aligned} m_A(t+1) &= \tanh\left(\frac{1}{2}(J_{AA}m_A(t) + J_{AE}m_E(t))\right) \\ m_E(t+1) &= \tanh(J_{EA}m_A(t)) \end{aligned} \tag{11}$$

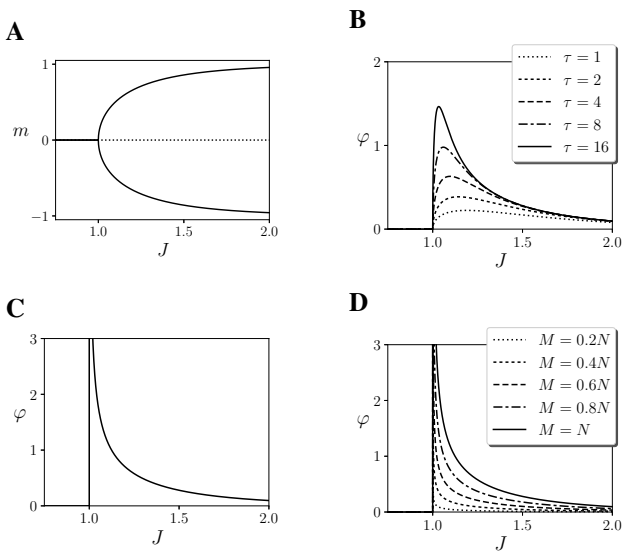

Figure 2: **(A)** Magnetization of the infinite size kinetic Ising model. **(B)** Value of $\varphi_{\mathcal{M}_N}(\tau)$ for different temporal spans. **(C)** Value of $\varphi_{\mathcal{M}_N}(\tau \to \infty)$ for an infinite temporal span. **(D)** Value of $\varphi_{\mathcal{M}_M}(\tau \to \infty)$ for different mechanisms of size M and an infinite temporal span.

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 an similar solution than the previous case, presenting a pitchfork bifurcation at a critical point (Figure 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 $s(t)$ (Figure 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:

$$\begin{aligned} \varphi_A &= \frac{1}{4}\Big(\Big|\sum_{\mathcal{R}=A,E} F'_{\substack{\mathcal{R} \\ MIP_A}}(m_0, \tau, 0)\Big|\Big) \\ \varphi_{AE} &= \frac{1}{4}\Big(\Big|\sum_{\mathcal{R}=A,E} F'_{\substack{\mathcal{R} \\ MIP_{AE}}}(m_0, \tau, 0)\Big|\Big) \end{aligned} \tag{12}$$

In Figure 3.C,E we estimate the value of φ_A, φ_{AE} for $\tau \to$

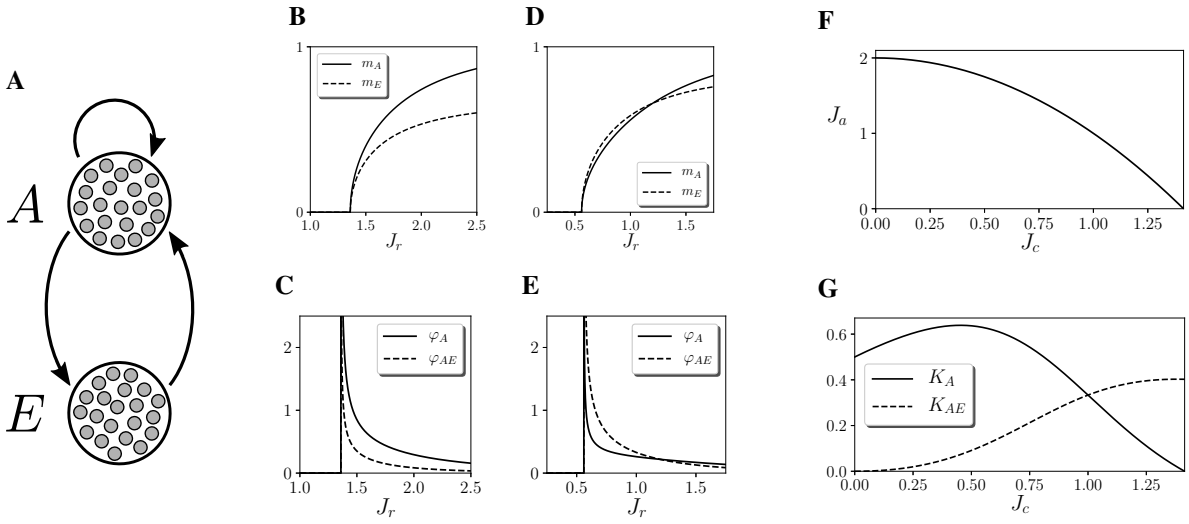

Figure 3: **(A)** Basic agent connected to an environment. **(B,C, D, E)** Values of the mean fields (only positive values are shown) of the stable solution (top) and $\varphi(\tau \to \infty)$ (bottom) for the agent and environment nodes of the model at stability for $J_c = 0.8$ (left) and $J_c = 1.2$ (right) and different values of J_r. **(F)** location 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 which is the most irreducible unit of the system.

∞ 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 results by solving Equation 11 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:

$$F_A(m_0, \tau \to \infty, 0) = \sqrt{\frac{3(J_{AA} + J_{AE}J_{EA} - 2)}{J_{AE}J_{EA}^3 + \frac{1}{4}(J_{AA} + J_{AE}J_{EA})^3}}$$
$$F_E(m_0, \tau \to \infty, 0) = \tanh(J_{EA}F_A(m_0, \tau \to \infty, 0)) \tag{13}$$

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 $J_{AA} + J_{AE}J_{EA} = 2$ (Figure 3.F). From here, we get:

$$F'^{A}_{MIP_A}(m_0, \tau, 0) = \frac{3}{2}J_{AA}\frac{1}{J_{AE}J_{EA}^3 + \frac{1}{4}(J_{AA} + J_{AE}J_{EA})^3}$$
$$\cdot(\frac{1}{4}(J_{AA} + J_{AE}J_{EA}) \cdot F_A(m_0, \tau \to \infty, 0) - \frac{1}{F_A(m_0, \tau \to \infty, 0)})$$
$$F'^{E}_{MIP_A}(m_0, \tau, 0) = \frac{J_{EA}}{\cosh(J_{EA}F_A(m_0, \tau \to \infty, 0))^2}F'^{A}_{MIP_A}(m_0, \tau, 0)$$
$$F'^{A}_{MIP_{AE}}(m_0, \tau, 0) = \frac{3}{2}J_{AE}J_{EE}\frac{1}{J_{AE}J_{EA}^3 + \frac{1}{4}(J_{AA} + J_{AE}J_{EA})^3}$$
$$\cdot(\frac{J_{EA}^2}{3} + \frac{1}{4}(J_{AA} + J_{AE}J_{EA}) \cdot F_A(m_0, \tau \to \infty, 0) - \frac{1}{F_A(m_0, \tau \to \infty, 0)})$$
$$F'^{E}_{MIP_{AE}}(m_0, \tau, 0) = \frac{J_{EA}}{\cosh(J_{EA}F_A(m_0, \tau \to \infty, 0))^2}$$
$$\cdot(F'^{A}_{MIP_{AE}}(m_0, \tau, 0) - F_A(m_0, \tau \to \infty, 0))$$

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

$$\begin{aligned} \varphi_A &= J_{AA}K(J_{AA} + J_{AE}J_{EA} - 2)^{-1/2}, \\ \varphi_{AE} &= J_{AE}J_{EA}K(J_{AA} + J_{AE}J_{EA} - 2)^{-1/2}, \\ K &= \frac{3(1 + J_{EA})}{\sqrt{J_{AE}J_{EA}^3 + \frac{1}{4}(J_{AA} + J_{AE}J_{EA})^3}} \end{aligned} \tag{14}$$

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 Figure 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.

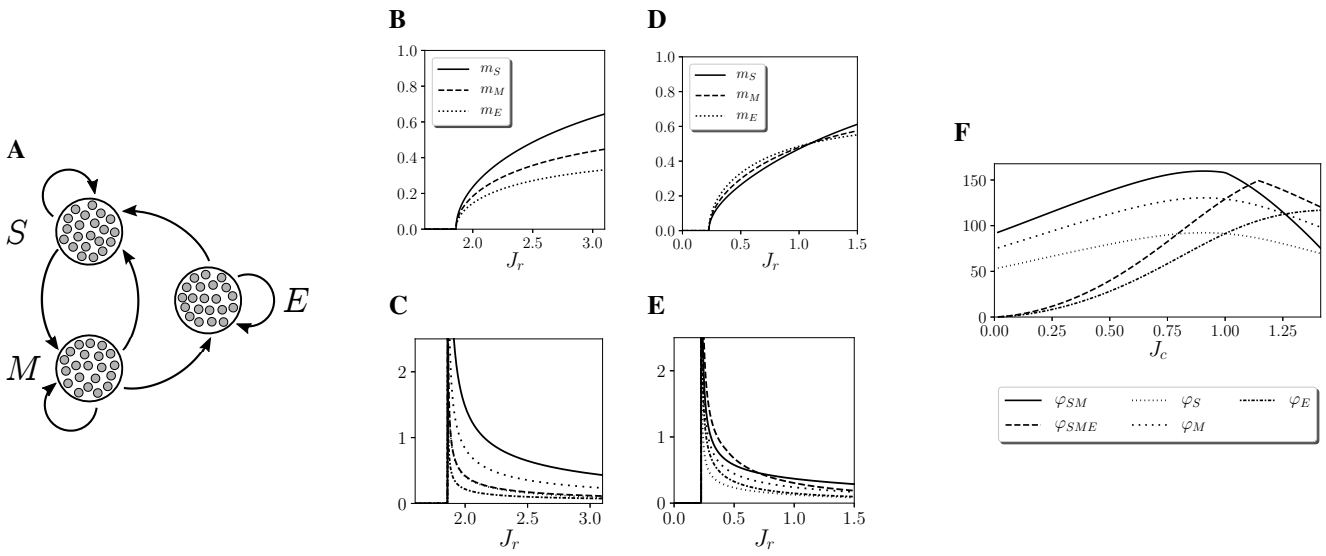

Figure 4: **(A)** Basic sensorimotor agent connected to an environment. **(B,C,D,E)** Values of the mean fields (only positive values are shown) of the stable solution (top) and $\varphi(\tau \to \infty)$ (bottom) for the agent and environment nodes of the model at stability for $J_c = 0.8$ (left) and $J_c = 1.2$ (right) and different values of J_r. **(F)** Approximated values of $\varphi(\tau \to \infty)$ near the critical point (approximated by $dx = 10^{-10}$), showing which is the most irreducible unit of the system.

Basic Sensorimotor Loop

We have just used integrated information for delimiting an agent interacting with a 'passive' environment showing no self-interaction. This is not a common scenario, since typically environments display their own dynamics. In order to portray a scenario with an agent regulating its interaction with an environment with 'active' dynamics, we model a 'sensorimotor agent' able to receive input from the environment in one region of its system, and affect the environment from the activity in other region (Figure 4.A).

$$\begin{aligned} m_S(t+1) \\ = \tanh(\frac{1}{3}(J_{SS}m_S(t) + J_{SM}m_M(t) + J_{SE}m_E(t))) \\ m_M(t+1) = \tanh(\frac{1}{2}(J_{MS}m_S(t) + J_{MM}m_M(t))) \\ m_E(t+1) = \tanh(\frac{1}{2}(J_{EM}m_M(t) + J_{EE}m_E(t))) \end{aligned} \tag{15}$$

A key aspect of autonomy is the ability of an agent to *modulate* the coupling with its environment (Barandiaran et al., 2009), generating an *interactional asymmetry* between agent and environment. We represent this by defining a basic structure of couplings $J_{SE} = J_{MS} = J_{EM} = J_c$ and self-couplings $J_{SS} = J_{MM} = J_{EE} = 1$. Finally, we add a recurrent connection in which the motor region feeds back into the sensor region $J_{SM} = J_r$. We also assume $r_S = r_M = r_E = 1/3$. We calculate φ as in previous cases, although here different candidates for MIP are tested (from all possible cuts affecting just one element) and the one minimizing integrated information is chosen.

For $J_c = 0.8$, we observe in Figures 4.C,E that the subsystem presenting a higher degree of integrated information corresponds to the agent SM, diverging at a critical point similar than in previous cases. However, if the coupling increases and $J_c = 1.2$, the agent is only the most integrated unit for large values of J_r, while the critical point of divergence shows that the most integrated unit is the whole agent-environment system SME.

Moreover, we approximate Equation 15 using the first term of the Taylor series of the hyperbolic tangents to find the position of the critical point, finding it at $Ja = (2 - Jc^3)/Jc$. Although in this case we cannot approximate the constant multiplying the diverging values of φ, we can approximate its values close to the critical point. For example, we approximate the value of $F'(m_0, \tau \to \infty, 0) \approx (F(m_0, \tau \to \infty, dx) - F(m_0, \tau \to \infty, 0))/dx$ for different partitions for a point close to the critical point, using a value of $dx = 10^{-10}$. In Figure 4.F, we observe a similar transition than in the previous case, from a case where φ_{SM} is dominant to a case where φ_{SME} is the principal integrated unit. Yet in this case, maybe due to a highest degree of agent-environment asymmetry, the region where φ_{SM} is dominant is slightly wider.

Discussion

We have presented a method to compute integrated information for infinite size mean field kinetic Ising models with quasi-homogeneous infinite-range connectivity. We have

shown how integrated information measures diverge when our models are near critical points. Furthermore, we have shown how integrated information can be used to effectively define the causal boundaries between a system and its environment. For doing so, some of the assumptions of current formulations of IIT had to be modified.

Our models, although highly idealized, allow us to speculate about some of the properties of autonomous organization and the nature of the transitions related to it. 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 near critical points the level of total integrated information diverges, suggesting that systems that are organized into coherent autonomous entities need to organize themselves close to critical points in their parameter space to maintain their level of integration as their size grows. This is relevant for some questions in origins of life research, such as why life appears as a jump in biochemical organization with no apparent intermediate steps.

Our results connect the intensity of this causal boundary with some phenomena related to criticality. Systems near critical points are maximally sensitive to changes in some directions of its parameter space (generally measured as the susceptibility of the system). Here, integrated information measures are captured by applying different partitions to the system which were interpreted as changes in particular directions of the parameter space of the system. Thus, the level of integrated information of the system corresponds to the susceptibility of the system for the minimum information partition, i.e., the partition with the less significant effect on the system causal powers. In the framework of IIT, systems highly sensitive to their minimum information partition are interpreted as maximally irreducible units. This connects ideas from IIT with properties that have been postulated as pervasive of living beings such as self-organized criticality. We can speculate that living autonomy consists in systems capable of self-organizing near points in which they can maintain maximal sensitivity to the integrity of their internal organization while they interact with the environment.

Acknowledgements

M.A. was supported by the UPV/EHU post-doctoral training program ESPDOC17/17 and project TIN2016-80347-R funded by the Spanish Ministry of Economy and Competitiveness.

References

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

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

Marshall, W., Kim, H., Walker, S. I., Tononi, G., and Albantakis, L. (2017). How causal analysis can reveal autonomy in models of biological systems. *Phil. Trans. R. Soc. A*, 375(2109):20160358.

Maturana, H. R. and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Boston Studies in the Philosophy and History of Science. Springer Netherlands.

Moreno, A. and Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. History, Philosophy and Theory of the Life Sciences. Springer Netherlands.

Oizumi, M., Albantakis, L., and 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., and Tsuchiya, N. (2016). Measuring Integrated Information from the Decoding Perspective. *PLOS Computational Biology*, 12(1):e1004654.

Walker, S. I. and Davies, P. C. W. (2013). The algorithmic origins of life. *Journal of The Royal Society Interface*, 10(79):20120869.

Appendix

Measures in this paper are inspired by the IIT framework, although we apply some modifications and simplifications. 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 $t+1$.

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.

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).

We simplify the calculation of the probabilites $p(\{s_i(t+\tau)\}_{i\in\mathcal{M}}|\{s_i(t)\}_{i\in\mathcal{M}})$ and $p^{cut}\{s_i(t+\tau)\}_{i\in\mathcal{M}}|\{s_i(t)\}_{i\in\mathcal{M}})$ by using a mean field approximation. In IIT, 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 removing the connection.

Finally, once φ is computed, IIT proposes a second level of calculations for computing Φ where new bidirecional partitions are applied to the system. In our case, given the homogeneity of the system applying a second level of partitions produces similar results and for simplicity we did not apply it.