Agency and Integrated Information in a Minimal Sensorimotor Model

Miguel Aguilera[1,2], Carlos Alquézar[1,3] and Manuel G. Bedia[1,3]

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

Abstract

The concept of agency is of fundamental importance for Cognitive Science. However, usual definitions of agency are loose and the work to capture and measure it using mathematical tools is still in its infancy. Recently, the framework of integrated information theory has been proposed to capture the causal boundaries of biological autonomous systems. Here, we test measures of integrated information theory in a minimal model to test its capacity to identify and delimit an autonomous agent interacting with an environment. Doing so, we reformulate some aspects of current definitions of agency using insights from integrated information in our models. Specifically, we propose a redefinition of how we capture the ability of an agent to modulate its interaction with the environment in terms of the control of the emergent causal structure of the agent-environment system. In this way, we propose an operational definition of agency based on the capacity of a system to modulate its causal boundary, extending and reducing it by functionally open and closing sensorimotor loops, and coupling the agent to different environmental processes. This allows us to formulate a tentative measure for our definition of agency and test it in minimal models of sensorimotor interaction, which we test in a minimal agent evolved to solve a simple task.

Introduction

The notion of agency is essential in fields as Artificial Intelligence and Cognitive Science. The need to define and clarify this concept is considered as one of the most crucial contributions capable of improving cognitive modelling practices. Although one finds a lot of definitions of agency in the literature (Wooldridge and Jennings, 1995; Russell and Norvig, 2016; Maes, 1993), when observed in detail, most of these definitions rely on intuitive notions and undefined terms. In particular, in engineering domains, it is very common to use a vague and uncritical use of this notion.

The difficulty of proposing a definition of agency increases when we seek a description that allows us to identify living systems with more or less clear physical boundaries but also sets of processes and, more generally, collective or cultural organizations. That is, defining agency becomes challenging when we intend to go beyond the standard notion of agent as a physical system with sensors and effectors. In practical terms, and in order to obtain a useful scientific definition it becomes necessary to provide an operational, quantitative and precise characterization of the object of study beyond an intuitive notion of agency.

Previous work has been oriented to outline a definition from these considerations. For example, Barandiaran et al. (2009) propose three different aspects of agency that constitutes a description of what an agent should be: (i) it is a distinguishable entity different from its environment (individuality), (ii) it is an active source of activity and interaction in its environment (asymmetry) and (iii) it is able to actively regulate their interactions according to some internal goals or norms (normativity). We take this definition of agency as starting point because we are interested in having an approximation for the agency in tune with the notion of embodied cognition. Thus, in order to define admissible conditions for agency, sensorimotor coupling and modulation of the interactions agent-world need to be considered. In this approach, an agent is understood not only as a structure that is individualized by itself, through autonomous mechanisms of organization, but it shows a sensorimotor dimension (agents should be able to maintain interaction and flexible sensorimotor coupling with the environment, Figure 1). Moreover, it explicitly involves a temporal dimension in the coordination dynamics between agent and environment (agents should be able to modulate the coupling in an adaptive manner).

In any case, we believe that this definition has some limitations that we intend to overcome. For example, the authors provide a generative definition (that is, a description of an organization capable of satisfying a set of requirements) but we seek an operational definition, a criterion that allows us to quantitatively evaluate the degree of agency of a system. On the other hand, a generative definition is necessarily sequential since it proposes a set of requirements in the form of a list. In (Barandiaran et al., 2009), the individuality condition is understood as a precondition for the modulation of the couplings with the environment. However, it is not clear that this occurs in natural living systems, where we could observe how the three mentioned conditions taking place si-

multaneously at a given instant of time.

Finally, other of the most controverted aspects of this contribution is how the modulation of the coupling is conceptualized. Once the agent and environment have been defined, it is proposed that the agent modulates its interaction by changing the value of a set of predefined conditions on the coupling. It is hypothesized, henceforth, that the agent can systematically and repeatedly modulates its structural coupling by controlling the value of certain constraints and that these changes typically are not induced by the environment. As well as we should avoid assuming that we know what the boundaries of an agent are before defining it, we should also avoid definitions that first tell us what an individual entity is, and then impose a specified interaction in terms of predetermined variables.

Instead, we propose that this modulation is a systemic effect of the agent-environment coupling and not the fine-tuning of certain variables by the agent. In this paper we introduce a definition referred to something that spreads throughout, system-wide, affecting the whole and not in terms of its elements.

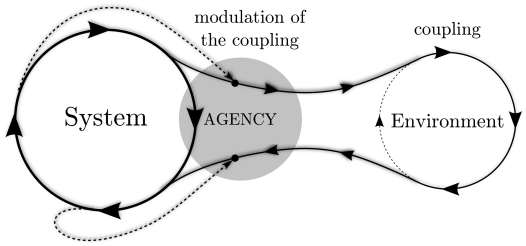

Figure 1: Illustration of an autonomous agent modulating its coupling with the environment.

To advance in this challenge, we take inspiration in integrated information theory (IIT, Oizumi et al., 2016). This theory provides an interesting approach to determine the individualization of certain processes that are integrated. It is, in other words, an operative measurement of irreducibility of a system based on the integration of processes and delimitation of causal boundaries. Moreover, IIT has been proposed to identify the integrated causal circuits that compose an autonomous biological systems (Marshall et al., 2017). Here, we propose to adapt some measures from IIT in order to advance towards an operational definition of agency. For doing so, we must use it in a fashion that allows us to take into account aspects of sensorimotor interaction, as we have proposed above, and this extension should be operationally determined without the need to fragment the two dimensions of the agency (individualization and regulation of its coupling). Achieving this, we also obtain a definition that meets theoretical conditions with operative requirements, providing a tentative measure to determine not only autonomous processes but a characteristic signature of agency.

For doing so, we postulate, as a proof of concept, a minimal model of a sensorimotor entity in interaction with an environment. We hypothesize that the agent's identity may emerge around a transition where the agent has the ability to intrinsically control the transit between modes of coupling and decoupling from its environment, and we define a criterion to capture this phenomena using current tools from integrated information theory. To further explore this idea, we consider another minimal agent designed to solve a non-trivial task requiring a high level of sensorimotor integration. We find that, when the task to solve is not trivial, agents able to successfully solve this task are poised near a similar transition in which the causal boundary of the integrated system goes back and forth from the agent to the whole agent-environment coupled system. Finally, we discuss the implications and possible generalization of our findings.

Integrated Information Theory and Individuality

Recent efforts have tried to quantify individuality and autonomy using information theory over the path of a system dynamics (Bertschinger et al., 2008; Krakauer et al., 2014). Still, these approaches presents some limits in order to distinguish a system from its environment. Typically, while nonlinear correlations of a dynamical system can be described in dynamical or information theoretical terms, they cannot be used to directly infer the boundary between an autonomous system and its environment.

Latterly, 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 stability, one could capture the causal forces integrating the behaviour of the system by observing it when some perturbations are imposed. Specifically, Marshall et al. (2017) have proposed the framework of integrated information theory (IIT, Oizumi et al., 2014).

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 by the integrated information φ of the subset of elements, which when larger than 0 indicates that the subset at its current state constraints the past and future states of the system in an irreducible way. By irreducibility it is understood that even the less disrupting bipartition of the system in two disconnected halves (that is called the minimum information partition, MIP) would imply a loss of information. Asides from computing integrated information at the level of mechanisms, IIT postulates a composite measure Φ, which is calculated from the set of all mechanisms (each one defined by a value of φ) obtained in the original system and the system under bidirectional partitions. A system with $\Phi > 0$ is described as

forming an 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. A more detailed description of the steps for calculating Φ is detailed in the Appendix.

Model

In order to advance towards an operational definition of agency, we propose to test our ideas in a very simple model. The model presents a general case of some elements of a system engaged in a loop of interaction. The question is whether we can delimit the boundaries of an agent to some of the elements when an asymmetry arises in the interaction of this subsystem and what is outside of it.

First, we postulate a minimal model defining causal temporal interactions among the elements that constitute it. Looking for generality, we use the least structured statistical model (i.e., a maximum caliber model, Pressé et al., 2013) establishing causal correlations between pairs of units from one time step to the next. We study a kinetic Ising model where N Ising elements s_i evolve in discrete time, with synchronous parallel dynamics. Given the configuration of units at time $t-1$, $s(t-1) = \{s_1(t-1), \ldots, s_N(t-1)\}$, the units $s_i(t)$ are independent random variables drawn from the distribution:

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

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 element and the couplings between pairs respectively, and β is the inverse temperature of the model. Without loss of generality, we can assume a $\beta = 1$.

Looking for a minimal example, we describe the case of three units S, M, E engaged in a loop of interaction (Figure 2.A). All elements have self-connections and each element influences the immediately posterior one $E \to S \to M \to E$ in a circular loop. In order to introduce an asymmetry in the interaction, we add an extra connection $M \to S$, with the objective of allowing the SM system to modulate the input received from an hypothetical environment E.

Individuality in a minimal sensorimotor model

In our system, we can easily apply IIT over its causal structure (Figure 2.B). We do so using the PyPhi toolbox (Mayner et al., 2017). IIT provides different values of Φ for different subsystems quantifying the level of integration of its relations. As an example, we apply IIT over a specific configuration of the system, where $J_{SS} = J_{MM} = J_{EE} = 0.25$, $J_{SE} = J_{MS} = J_{EM} = 1$, $H_S = H_M = H_E = 0$, and

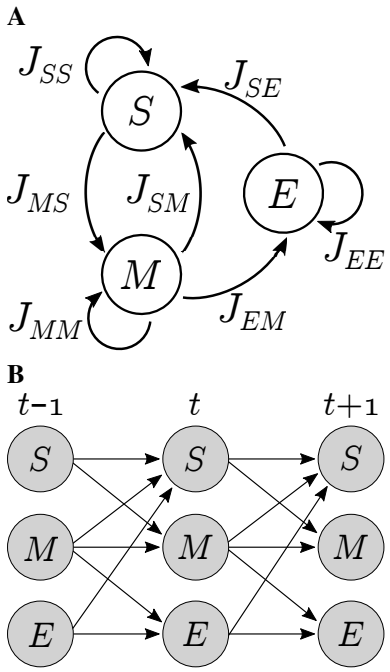

Figure 2: Description of the model. **(A)** The structure of the kinetic Ising model consisting of three elements (S, M and E). **(B)** The system's causal structure of dependencies with future and past states.

J_{SM} is a free parameter that determines the strength of the reentrant connection modulating the input of the system.

In this simple system, we find two subsystems with a value of Φ larger than zero: the one formed by the units SM and the one comprised by the whole system SME. For different values of J_{SM}, the results are shown in Figure 3.A.

According to Marshall et al. (2017), IIT can identify causal boundaries defined as subsets of elements that define maximum local values of intrinsic and irreducible cause-effect relations. In that sense, we estimate that Φ can be a good indicator of the level of individuality of a system, although some extra steps would be necessary for this individuality to constitute an autonomous agent.

We can observe the levels of Φ in our simple model in Figure 3.A, where we show the mean values of Φ_{SM} and Φ_{SME} as well as the area comprised by their maximum and minimum values. For large values of J_{SM}, the value of Φ_{SM} increases, indicating that the coupling SM defines an emergent causal boundary that separates it from the environment E. If we take larger values of J_{SM}, the system SM is practically 'blind' to its environment (since the input from M to S has a much larger influence). Similarly, in cases where J_{SM} is very small, the boundary around SM disappears and we can identify an individuality at the level of the whole system SME with a large value of Φ_{SME}. In this case, all the elements are closely interacting, and we can not find an

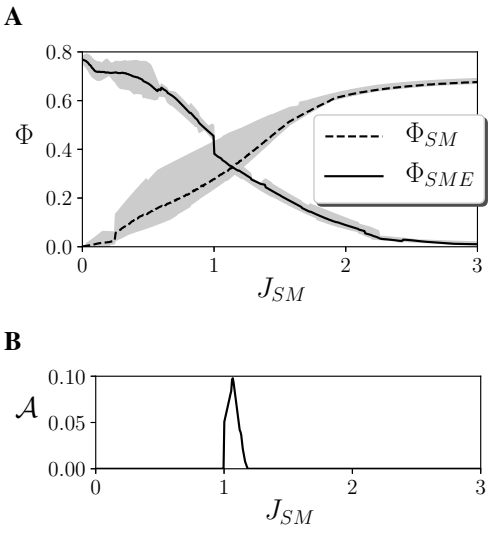

Figure 3: Effect of the parameter J_{SM} in the integration of information of the system. **(A)** Comparison of the mean Φ values obtained in both the whole system Φ_{SME} (continuous line) and the subsystem Φ. The gray area represents the interval between the minimum and maximum Φ values. **(B)** Proposed measure of agency $\mathcal{A} = \langle|\Delta\Phi|\rangle - |\langle\Delta\Phi\rangle|$, capturing fluctuations between Φ_{SM} and Φ_{SME} as local maxima of Φ.

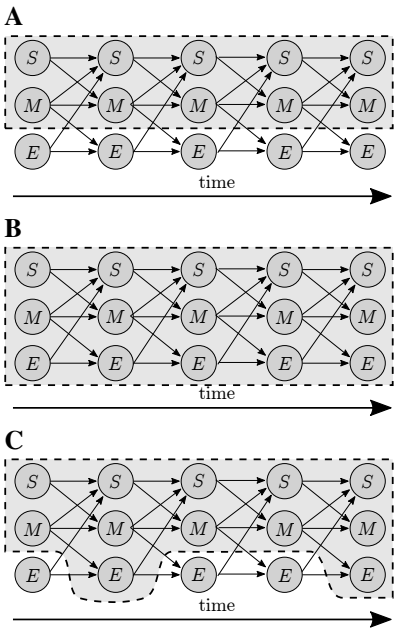

Figure 4: Three scenarios of the evolution of the causal boundaries of an agent-environment system. (**A**) The most integrated unit is the agent ($\Phi_{SM} > \Phi_{SME}$). (**B**) The most integrated unit is the joint agent-environment system ($\Phi_{SM} < \Phi_{SME}$). (**C**) The most integrated unit fluctuates over time between the previous cases.

asymmetry in the relations between elements that can define an isolated agent.

Somehow, the most interesting situation appears in the case where there is an uncertainty about which subsystem constitutes an individuality. For values roughly around $J_{SM} = 1.1$ there is a situation in which Φ_{SM} could be either higher or lower than Φ_{SME} (note that Φ is state dependent and its value changes in time). We can define this uncertainty through the variable $\Delta\Phi = \Phi_{SM} - \Phi_{SME}$. In some cases, e.g. around $J_{SM} = 1.1$, the span of $\Delta\Phi$ will be high enough that the local maxima of Φ will shift back and forth from the agent SM to the agent-environment system SME. In this case, if we take a local maximum of Φ to be the main causal structure of a system at a specific moment, we can interpret shifts in Φ as an agent-environment sensorimotor loop that can be opened and closed at different moments of time. We hypothesize that this phenomenon is a good candidate to describe the ability of an agent to modulate its sensorimotor coupling.

As introduced above, the main requirements for autonomy are: (i) the constitution of an agent as an individuated unit separated from its environment and (ii) the emergence of an agent-environment asymmetry in which the agent actively modulates its interaction with the environment. Establishing condition (i) as a prerequisite for testing (ii) is problematic because, as we have just seen, defining the boundary between agent and environment is not easy, and in the most interesting cases this boundary is going to extend back and forth covering some elements of the environment as the agent is engaged in sensorimotor coupling.

Based on these intuitions, we formulate an alternative approach for an operational definition of agency. We propose that agents are entities coupled to external environments capable of generating emergent causal boundaries that delimit the sensorimotor integration at a particular moment. The limits of this causal boundary will extend and contract with time, as the internal mechanism of the agents couple and decouple from different sensorimotor loops. Agency emerges precisely throughout open loops that are formed in the course of interactions, extending the boundary of causal integration of an agent to elements of the environment at different moments of time.

Following this idea, we formulate a tentative measure of agency $\mathcal{A} = \langle|\Delta\Phi|\rangle - |\langle\Delta\Phi\rangle|$. This measure tries to assess the changes in the role of leading mechanism that constitutes the identity in the system (i.e. the location of the subsystem with maximum Φ). A value of $\mathcal{A} = 0$ would indicate that always the same subsystem has the higher level of integration, while values upper zero would indicate that the location of the mechanism with higher integration changes with time. Thus, in the latter cases we could define an integrated dy-

namic core that describes the causal structure of an agent as a dynamical entity that open and closes different loops of interaction.

Analyzing the value of $\mathcal{A}$ across systems with different values of the parameter J_{SM}, we get the results shown in Figure 3.B. There, it is illustrated how the value of $\mathcal{A}$ arises at a narrow range close to $J_{SM} = 1.1$. Taking a look at this point, we could consider it as a transition point between two regimes in the parameter space where, in one side, always the corresponding subsystem of the agent arises as the most integrated one (Figure 4.A) and in other regime, always the whole system is the maximally integrated unit (Figure 4.B). But, in the transition region, the limits of the most integrated unit of the system change depending on the state of the system (Figure 4.C). One possible interpretation of these fluctuations of the causal boundary could be to view it as an scenario in which an 'agent' is constituted throughout the open loops that can engage in different modes of coupling with its surrounding environment.

Interactional asymmetry in a minimal agency task

In this part we apply the proposed measure in terms of integrated information to a minimal model of an agent engaged in a non-trivial task. We design an agent performing a task in an environment that can be interpreted in cognitive terms, while maintaining the same statistical structure than the previous minimal model.

The environment consists of a binary world composed by two squares (Figure 5), where it is only possible to move between two positions: left or right (i.e. $s_E = \pm 1$). As in the previous case, we consider an agent composed of a sensor and motor units (S and M) able to perceive and move in this environment. Depending on the location of the agent in the environment, the sensor unit S perceives light or darkness (i.e. $s_S = \pm 1$). The position of the lights changes randomly, with a probability of change $P_{change} = 2^{-5}$ at each time step. The goal of the agent is to maximize the time it spends in the illuminated square.

The direct connections (J_{EM}, J_{SE}) related to how the agent and environment influence each other are fixed. The influence from the motor to the environment is set to $J_{EM} = J_c$. The local field and self-connection of the environment are set to $H_E = 0$ and $J_{EE} = 0$. The position of the light will determine the influence of the environment to the agent. When the right square is illuminated, $J_{SE} = J_c$. On the other hand, when the left square is illuminated $J_{SE} = -J_c$. The rest of the connections affecting S and M, i.e., $H_S, H_M, J_{SS}, J_{MM}, J_{SM}, J_{MS}$ will be tuned for maximizing the fitness of the agent withing the range $[0, 5]$.

A fitness function is designed to select agents that are able to perform well for both possible environments (left and right light). For computing the fitness value, agents are simulated for 100 trials of duration 500 steps starting from a random state, for the two scenarios (light either at the left or right square). Then, the fitness value is defined as $F = \sqrt{\langle L \rangle_{left} \langle L \rangle_{right}}$, where $L(t) = 1$ when the position of the agent E corresponds with the illuminated square, and $L(t) = 0$ otherwise.

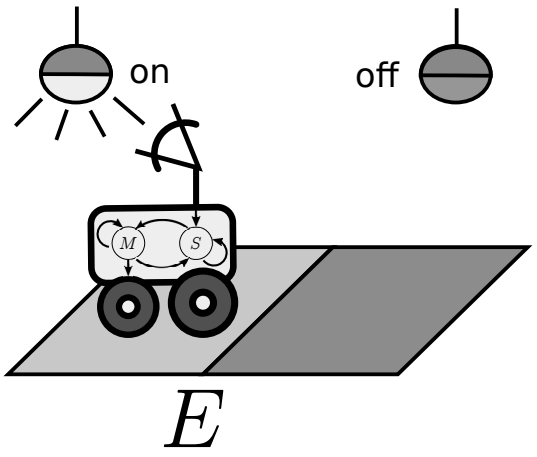

Figure 5: Illustration of the task. A sensorimotor agent must maximize its exposition to light in a noisy environment.

For 36 values of J_c in the range $[0.1, 3.5]$, we run a microbial genetic algorithm (Harvey (2009)) in order to obtain the agent with the highest fitness. The genetic algorithm simulates a population of 100 agents during 5000 generations. Recombination and mutation rates are set to 0.5 and 0.1 respectively. In Figure 6.A it is illustrated the evolution of the fitness related to the value of the connections with the environment. Notice that fitness increases when the connections become stronger. The interpretation of this correlation would be that, for higher values of J_c, the task becomes easier for the agent, because the interactions with the environment are dominating the internal dynamics, so the agent can just interact with the environment in a reactive fashion.

Once an agent has been evolved, in order to analyze its relations with the environment in terms of integration of information, we simulate the system during $T = 10000$ steps and, for each state obtained at each time, we calculate the corresponding value of Φ and $\Delta\Phi$ associated to that state. Doing this, we get the temporal evolution of the integration of information for both the agent and the whole system. For each agent, we get a single value of agency $\mathcal{A}$ that determines the level of how the agent integrates and disintegrates the environment with itself across time.

Making a comparison over systems with different values of J_c, we analyze the level of agency of the 20 agents with the best performance resolving the task (Figure 6.B). We find that, for most values of J_c, most agents present $\mathcal{A} = 0$, suggesting that a level of agency is not necessary to obtain the maximum fitness available for an agent. We find an exception around $J_c = 1$, where most agents have a value of $\mathcal{A}$

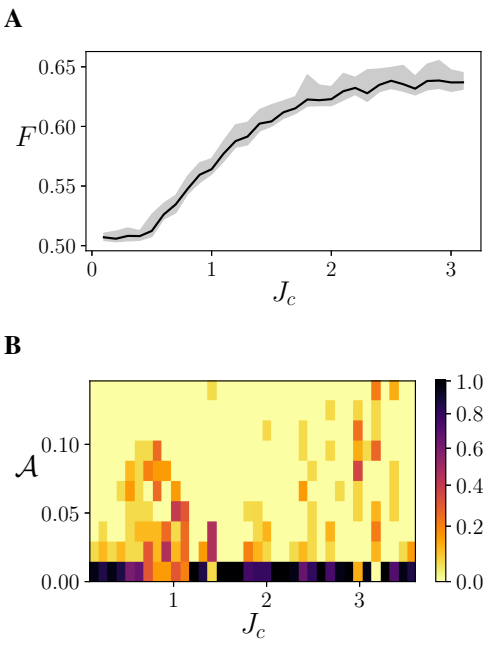

Figure 6: Results of applying the measure of agency across different setups. (**A**) Evaluation of the task. In the Figure is shown the fitness function of the best agent performing the task across each value of the parameter J_c. (**B**) Histogram of the values of $\mathcal{A}$ for the best 20 agents obtained with the genetic algorithm for each value of J_c.

larger than zero. As well, for larger values of J_c some agents present large values of $\mathcal{A}$, although the majority presents values of zero. Although further tests are necessary, we speculate that precisely the region around $J_c = 1$ is the region where solving the task is possible (the signal to noise ratio starts to be significative) but not trivial (the agent must integrate information about its input to filter out input noise). In this scenario, it may be difficult to solve the task without some level of agent-environment asymmetry, since the agent must integrate a noisy imput signal with limited internal resources. As for the large values of $\mathcal{A}$ that we find for large values of J_c, we hypothesize that when the task is simpler to solve, a larger diversity of structures may be able to achieve higher fitness, thus presenting some agents with high $\mathcal{A}$, although this is not necessary to solve the task.

Discussion

In this paper we have proposed an operational definition of agency based on ideas from integrated information theory that offers a framework for measuring the level of integration of the causal structure of subsets of elements of a system. Specifically, inspired by previous work by Barandiaran et al. (2009), we have proposed that a definition of agency should be able to capture at the same time the ability of an

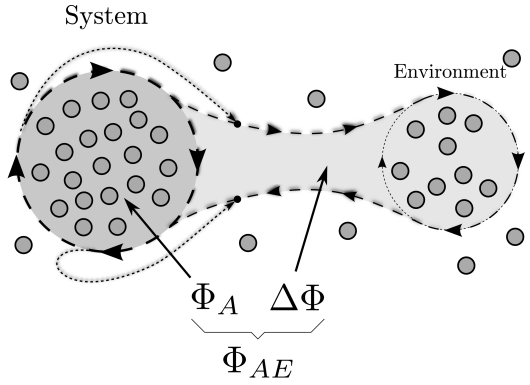

Figure 7: Illustration of a situation when we can not differentiate an agent of its environment

agent to constitute an integrated whole (individuality) and to modulate its coupling with the environment to extend its boundary of integration in order to incorporate elements of the environment at certain moments of time (interactional asymmetry, Figure 7).

Moreover, we have tried to go beyond a generative definition such as that proposed by the authors, i. e., a set of necessary and sufficient conditions for a minimum conception of the agency, because it is not as useful as an operational definition that allows to quantitatively determine the degree of agency of a system. Our way of characterizing the agency offers the following properties: (i) it is an 'operative criteria', i.e. it specifies how to calculate and how to be applied in experimental domains being able to characterize the degree of agency of a system (and not only providing a list of requirements); (ii) it assumes the condition of 'sensorimotor agency' (it differs from internalist perspectives that understand the agency based on internal architecture of controllers but not at the embodied mechanisms of relations with the environment), (iii) it has a temporal dimension (it highlights the dynamic nature of the interactive regulation processes) (iv) it is conceived from a holistic perspective (avoiding the simplicity of sequential approaches in which the agent is first identified and then examined about how it interacts with the world).

We have tested this idea in a simple model evolved to perform an easy but non-trivial task. Using a minimal model of an agent and an environment, we have shown how there exist situations where the location of the maximally integrated structure of the system is not fixed but changes with time. At some moments, it only comprises the agent, while at others it is composed of the agent plus the environment it is coupled to. Throughout the paper, we have shown some evidence that our definition is a good identifier for this type of organizations capable of adaptively regulating its coupling with the environment.

Contrarily to Barandiaran et al. (2009), our approach does

not need to add *ad hoc* variables to describe how an agent might modulate the interaction with its environment. Instead, this modulation is described at an emergent level of how integrated information expands or shrinks to cover parts of this environment. Moreover, although our tests are implemented over very simple agents, the objective is to illustrate the kind of phenomena we could encounter in more complex systems.

As a further observation, we would like to remark that our proposal is in tune with works that point out that many of the difficulties for an adequate definition of agency are related to the fact that operational closure requires systems that constitute itself as unified wholes that can be regarded as separated from the environment although in continuous interaction with it. For example, contributions in extended cognition (Dotov et al., 2010) that analyze situations in which a subject and a tool constitute an extended device during smooth coping, which can be temporarily interrupted and again self-assembled during an action. Or works as (Fuchs, 2011), where the brain is conceived as a plastic system of open loops that are formed in the process of interaction with the environment and are closed to full functional cycles in each interaction.

Although further experimental tests are needed, we hope that this contribution could be a step in fields as autonomous robotics or artificial life towards the development of quantifiable artificial forms of agency, focusing on the question of how the emergence of sensorimotor loops relate to the autonomous constitution of a system.

Acknowledgements

MA and CA were funded in part by project TIN2016-80347-R from the Spanish Ministry of Economy and Competitiveness. MA was also supported by the University of the Basque Country UPV/EHU post-doctoral training program ESPDOC17/17. CA wishes to acknowledge funding from the Okinawa Institute of Science and Technology and the Initiative for a Synthesis in Studies of Awareness to assist to the ISSA Summer School 2017.

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.

Dotov, D. G., Nie, L., and Chemero, A. (2010). A Demonstration of the Transition from Ready-to-Hand to Unready-to-Hand. *PLOS ONE*, 5(3):e9433.

Fuchs, T. (2011). The Brain–A Mediating Organ. *Journal of Consciousness Studies*, 18(7-8):196–221.

Harvey, I. (2009). The Microbial Genetic Algorithm. In *Advances in Artificial Life. Darwin Meets von Neumann*, Lecture Notes in Computer Science, pages 126–133. Springer, Berlin, Heidelberg.

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

Maes, P. (1993). Modeling Adaptive Autonomous Agents. *Artif. Life*, 1(1-2):135–162.

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.

Mayner, W. G. P., Marshall, W., Albantakis, L., Findlay, G., Marchman, R., and Tononi, G. (2017). PyPhi: A toolbox for integrated information theory. *arXiv:1712.09644 [cs, q-bio]*. arXiv: 1712.09644.

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.

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

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Wooldridge, M. and Jennings, N. R. (1995). Intelligent agents: theory and practice. *The Knowledge Engineering Review*, 10(2):115–152.

Appendix

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. Given Equation 1, the computation of p is straightforward. 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 $\{s_i(t+1)\}_{i\in\mathcal{P}}$ (Figure 2.B). We describe the cause and effect repertoires of the system by the probability distributions $p_{cause}(\mathcal{P}_{t-1}|\mathcal{M}_t) = p(\{s_i(t-1)\}_{i\in\mathcal{P}}|\{s_i(t)\}_{i\in\mathcal{M}})$ and $p_{effect}(\mathcal{P}_{t+1}|\mathcal{M}_t) = p(\{s_i(t+1)\}_{i\in\mathcal{P}}|\{s_i(t)\}_{i\in\mathcal{M}})$.

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,

$$\begin{aligned} \varphi_{cause} = \\ \max_{\mathcal{P}} \left(\min_{cut} \left(D(p_{cause}(\mathcal{P}_{t-1}|\mathcal{M}_t), p^{cut}_{cause}(\mathcal{P}_{t-1}|\mathcal{M}_t)) \right) \right) \\ \varphi_{effect} = \\ \max_{\mathcal{P}} \left(\min_{cut} \left(D(p_{effect}(\mathcal{P}_{t+1}|\mathcal{M}_t), p^{cut}_{cause}(\mathcal{P}_{t+1}|\mathcal{M}_t)) \right) \right) \end{aligned} \tag{3}$$

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

$$\begin{aligned} cut = \{\mathcal{M}_1, \mathcal{P}_1, \mathcal{M}_2, \mathcal{P}_2\} \\ p^{cut}(\mathcal{P}|\mathcal{M}) = p(\mathcal{P}_1|\mathcal{M}_1) \otimes p(\mathcal{P}_2|\mathcal{M}_2) \end{aligned} \tag{4}$$

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

$$\varphi = min(\varphi_{cause}, \varphi_{effect}) \tag{5}$$

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:

$$\Phi = \min_{cut} D(C, C^{cut}) \tag{6}$$

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 movers 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. All computations in this paper were performed by the PyPhi software package (Mayner et al., 2017). Further details of the steps described here can be found in (Oizumi et al., 2014).