Adaptive **Behavior**

Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action

Xabier E. Barandiaran[1], Ezequiel Di Paolo[1], Marieke Rohde[2]
[1] *Center for Computational Neuroscience and Robotics & Department of Informatics, University of Sussex, Brighton, UK*
[2] *Multisensory Perception and Action Group, Max Planck Institute for Biological Cybernetics, Tübingen, Germany*

The concept of agency is of crucial importance in cognitive science and artificial intelligence, and it is often used as an intuitive and rather uncontroversial term, in contrast to more abstract and theoretically heavily weighted terms such as *intentionality*, *rationality*, or *mind*. However, most of the available definitions of agency are too loose or unspecific to allow for a progressive scientific research program. They implicitly and unproblematically assume the features that characterize agents, thus obscuring the full potential and challenge of modeling agency. We identify three conditions that a system must meet in order to be considered as a genuine agent: (a) a system must define its own *individuality*, (b) it must be the active source of activity in its environment (*interactional asymmetry*), and (c) it must regulate this activity in relation to certain norms (*normativity*). We find that even minimal forms of proto-cellular systems can already provide a paradigmatic example of genuine agency. By abstracting away some specific details of minimal models of living agency we define the kind of organization that is capable of meeting the required conditions for agency (which is not restricted to living organisms). On this basis, we define agency as an autonomous organization that adaptively regulates its coupling with its environment and contributes to sustaining itself as a consequence. We find that spatiality and temporality are the two fundamental domains in which agency spans at different scales. We conclude by giving an outlook for the road that lies ahead in the pursuit of understanding, modeling, and synthesizing agents.

Keywords agency · individuality · interactional asymmetry · normativity · spatiality · temporality

1 Agency as a Departure Point

The concept of agency plays a central role in contemporary cognitive science as a conceptual currency across different sub-disciplines (especially in embodied, situated, and dynamical approaches—Beer, 1995, 2003; Brooks, 1991; Pfeifer & Scheier, 1999). It owes this central role to its capacity to capture the notion of a behaving system while avoiding the endless discussions around alternative foundational terms such as *representations*, *intentions*, *cognitive subject*, *conscious being*, or *mind*. While an insect-like robot already seems to be a minimal instance of agency, the concept is open enough to also cover humans or even collective organizations. From the departure point of agency it is possible to envision a research program that pro-

Correspondence to: Dr. Xabier E. Barandiaran, Department of Informatics, University of Sussex, Falmer, Brighton, BN1 9QH, UK. *E-mail*: xabier.academic@barandiaran.net. *Tel.*: +44 1273 678062, *Fax*: +44 1273 877873

(2009), Vol 17(5): 367–386.
DOI: 10.1177/1059712309343819
Figure 1 appears in color online: http://adb.sagepub.com

ceeds from the bottom up, from the simplest embodied behavior, grounding higher level phenomena on increasingly complex forms of situated interactions and their underlying mechanisms. This program would, furthermore, be amenable to dynamical systems' modeling cutting across brain, body, and world and integrating different levels of mechanistic organization into the same explanatory framework. This possibility has generated considerable enthusiasm and has come to renew some of the foundations of cognitive science (Barandiaran, 2008; Beer, 1995; Christensen, 1999; Di Paolo, 2005; Hendriks-Jansen, 1996; Wheeler, 2005).

Yet, while the need to explicitly define *agency* has been recognized, most current researchers assume an intuitive and unproblematic notion of agency. As a large part of the literature shows, most researchers do not pay much attention to what it is that constitutes a system as an agent. Is a Khepera robot an agent, independently of its control architecture or its body, just in virtue of its capacity to move around an arena? What about a protocellular system pumping ions outside its membrane? Is a bird gliding on wind currents an agent? Do the tremors affecting a Parkinson disease patient count as agency? How can we justify the negation or attribution of agency to the above cases in a manner amenable to scientific scrutiny?

Despite the difficulty in providing a clear and precise answer to these questions, a loose or metaphorical concept of agency has helped to reconceptualize cognitive systems as inherently situated while grounding intelligent capacities on behavior-generating mechanisms (as opposed to abstract symbolic algorithms). However, it is time to move forward and to propose a deeper definition of agency, capable of addressing some fundamental issues that could bring natural agency to scientific scrutiny and improve our modeling practices (including the creation of artificial agents).

This article is organized as follows. First, we identify three key properties of agency: individuality, asymmetry, and normativity and argue why these are good candidates for necessary and sufficient conditions for agency. Next, we examine how these properties are already present in simple living systems (as highlighted by a long tradition in philosophical and theoretical biology and cognitive science). We shall then describe a minimal-template organization that meets those conditions in the sense that it generates phenomena that satisfy them (we shall call this a "generative definition"). We then discuss how spatiality and temporality are linked to and co-emerge with agency. We finally conclude with an evaluation of ongoing research in the modeling and synthesis of artificial agents.

2 What Agency Requires: Individuality, Asymmetry, and Normativity

In making a scientific definition we must balance two constraints in tension: we must be able to capture the meaning of the term as used intuitively in science and in everyday life and, at the same time, we must provide an operational and precise characterization. Everyday concepts can be slippery (yet perfectly useful), whereas strict definitions can easily become too narrow or too broad. Undoubtedly, we face this tension in the present attempt to define agency and we will evaluate our definition along these requirements. We recognize that properly characterizing such a complex and polysemic concept is not divorced from our attempts to study agency. Consequently, our theoretical understanding should not only depend on intuitive ideas and full-fledged preexisting theoretical accounts, but also on present and future empirical or modeling results. In other words, our definition is a proposal that may have to be revised after an empirical evaluation of its practical and theoretical implications.

The rise of synthetic robotic approaches in cognitive science and adaptive behavior modeling in the 1990s has led to an explosion of proposed definitions of agency (Wooldridge & Jennings, 1995). For instance, Russell and Norvig in their classical AI handbook (1995, p. 33) propose that "an agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors." Maes (1994, p. 136), on the other hand defines an agent as "a system that tries to fulfill a set of goals in a complex, dynamic environment"; Beer (1995, p. 173) considers an agent "any embodied system [that pursues] internal or external goals by its own actions while in continuous long-term interaction with the environment in which it is situated," while Smithers (1995, p. 97) states that "agent systems are systems that can initiate, sustain, and maintain an ongoing and continuous interaction with their environment as an essential part of their normal functioning." After an extensive review

of different definitions of agency (including some of those previously mentioned), Franklin and Graesser (1996, p. 25) conclude that "an autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future." Kauffman (2000, p. 8) has defined an agent as a system that "can act on its own behalf in an environment." Following his work, Ruiz-Mirazo and Moreno (2000) defend that minimal autonomous agents are those chemical systems capable of actively constraining their boundary conditions for self-maintenance. In a parallel manner, Christensen and Hooker (2000) state that "[a]gents are entities which engage in normatively constrained, goal-directed, interaction with their environment" (p. 133).

Most of these definitions rely strongly on intuitive notions of sensing, perception, action, goal, and so forth. This means that if we want to establish whether a given system is an agent in the first place, not only study its behavior, we have to clarify these intuitive terms. In this sense, many definitions are not complete: they rely on additional undefined terms. For instance, understanding agency and understanding action are parallel endeavors, and we should not presuppose one in order to define the other.

In order to proceed systematically toward a definition we shall first attempt a non-controversial description of what an agent is. Abstracting away from the particularities of the above definitions we can generalize that agency involves, at least, *a system doing something by itself according to certain goals or norms within a specific environment*.

From this description, three different, though interrelated, aspects of agency follow immediately: (a) there is a system as a *distinguishable entity* that is different from its environment, (b) this system is *doing* something by itself in that environment, and (c) it does so according to a certain goal or *norm*. A generative definition of agency has to account, at least, for these three requirements. Let us investigate them in more detail.

2.1 Individuality

First of all, in order for a system to be an agent, there must be a distinction between the system and its environment. This we shall call the *individuality* condition. The identity of an agent as an individual distinguishable from its environment is often taken for granted or seen as trivially irrelevant. Any characterization of agency is then limited to the establishment of the kind of relationship (representational, informational, intentional, adaptive, etc.) between a pre-given agent and its world. However, neither a specific environment nor agentive relations with this environment can exist without the constitution of an agent as an individuated system. This constitution is not just a precondition for agency, a separate issue that, once it happens, can be taken for granted so as to focus attention on the relation between system and environment. As we will argue, the interactive dimension of agency appears tightly coupled to the very constitution of its individuality. (We will use the terms individuality and identity interchangeably.)

When describing an undifferentiated process (e.g., a homogeneous hectometer of gas), it is impossible to talk about a system and its environment. There need to be distinguishable and relatively stable components or ensembles. However, even when faced with such components, the question of which of them are assigned to the system and which to the environment remains. As observers, we can in principle establish any arbitrary separation in this kind of situation. The problem of individuality becomes the problem of justifying which one we choose among the large set of possible and arbitrary distinctions between system and environment and why does the system qualify as an individual (and not just as a mere collection or arbitrary aggregate).

In some cases, the tendency is to justify such distinctions through the functionality (including the epistemic convenience) that the composite system holds for the observer. For example, we might agree on declaring that the table, pen, paper, computer, and lamp constitute the "workplace" system. But, in this case, no intrinsic force or process is lumping the components together, nor has the system as a whole (independently of us) a specific way of functioning and demarcating itself from the rest of the office. In other cases, a force might be lumping components together, like the strong forces at the nucleus of an atom or gravitational forces in planetary systems. Similarly, some kind of structural linkage could also do the job, as in the case of towing hooks in a clock. And yet, what belongs to the clock "as a clock" or to planetary systems remains unclear without an observer introducing an arbitrary criteria on the definition of the system in terms of the function it

sub-serves for her (such as measuring time) or labeling a composite as a system by mere convention (such as astronomic ensembles). Robots are often described as agents in these two senses only: because its mechanical structure is lumping the material components into a unitary motile entity and because the robot as a whole operates according to some performance criteria that the observer or designer judges useful or coherent.

In his essay on the *Biological Foundations of Individuality*, Hans Jonas describes a similar situation. He inquires about the nature of organic identity and this, he argues, cannot be the same kind of identity granted to artifacts: "the decisive observation, of course, is that to the artifact the identity is accorded; and, insofar as this requires a continuity of memory and tradition in those who do accord it, the identity is the function of another identity, namely, that established in memory, individual and social. This originative identity of the cognitive subject is a prerequisite for the accorded identity of the object. But this original identity, being that of living systems, is just a case of what we are inquiring into, … This [identity] we have acknowledged as owned by, not loaned to its subject." (Jonas, 1968, pp. 239–240).

A concept of agency that cannot account for the way in which an agent defines itself as an individual requires another agent (the observer) to perform the system–environment distinction. If we then have to justify the identity of this observer agent by means of another one, and so on, we enter an infinite explanatory regress. In contrast, an entity capable of distinguishing itself as an individual in the absence of an observer, as Jonas proposes for the case of living organisms, does not suffer from this problem.[1] We then note that, as opposed to other systems, agents appear as unified in themselves and do not depend on their being useful for an external entity or accorded on their identity by a community of other agents in order to become what they are. Therefore, *the first condition for the appearance of agency is the presence of a system capable of defining its own identity as an individual and thus distinguishing itself from its surroundings; in doing so, it defines an environment in which it carries out its actions.* Moreover, agents define themselves as individuals as an ongoing endeavor and through the actions they generate, a point to which we will return later on. This brings us to the next condition for agency.

2.2 Interactional Asymmetry

Once an individual is in place, exchanges of matter and energy are inevitable at some level; the system is coupled to its environment. However, the concept of agency is intuitively associated with that of action, not mere system–environment coupling or exchange. An agent is a system that *does* something as opposed to other natural entities to which we attribute no specific actions except metaphorically (e.g., "The sun rises"). In other words, an agent is a source of activity, not merely a passive sufferer of the effects of external forces. Similarly, an agent is not driven to act by internal, sub-systemic modules which subordinate the system to the triggering or isolated functioning of a local mechanism. In a sense yet to be properly disclosed, an agent as a whole drives itself, breaking the symmetry of its coupling with the environment so as to modulate it from within. We call this condition *interactional asymmetry*.

In order to understand this condition, we should proceed in stages. A first approximation demands that we look for this asymmetry in the causal structure of the interaction between an agent and its environment. It seems intuitively right that the agent causes its own actions and that this causal role is sufficient to describe the asymmetry we are after. Needless to say, the concept of causality in complex systems is problematic. In this context, we reduce it to two possible scenarios: an energetic and a statistical sense.

One way to understand interactional asymmetry in terms of the causal origin of action events is to consider, as others have done, an agent as responsible for managing and gathering the energy resources for action. For this line of thinking, the asymmetry requirement is expressed in terms of the capacity of the system to constrain energy flows to sustain coordinated processes that are in turn reused by the system in a circular manner (Kauffman, 2000; Ruiz-Mirazo & Moreno, 2000). It is argued that cellular organization, by coupling endergonic and exergonic reactions and channeling energy flows, can produce work, moving the behavior of the system away from its thermodynamic tendency. Such energy-based conceptions of cause match an intuitive notion of action and agency: the system is the energetic drive of an otherwise neutral or spontaneous coupling with its environment (*actively pumping* ions or *performing* chemotaxis as opposed to the *passive suffering* of an osmotic burst or *being moved* by currents or local fluctuations in a

pond). However, being a source of activity does not imply trying to constantly avert the effect of environmental forces through the investment of internally channeled energy, but often, on the contrary, being able to "surf" these effects in a specific direction.

Consider a bird gliding. Being pushed by the wind is not usually considered an action, and, in the above sense, it does not involve the investment of its own energy resources. However, birds usually let themselves be carried by the wind exerting control only by means of minimal movements; they exploit and use external forces rather than to counter them head-on. Is this not an action? A notion of causation in terms of the required energy investment could account for actions only in those cases where the energy is fully or primarily recruited by the system. Otherwise, it is not possible to say that there is an asymmetry in the interaction. Actions such as gliding become ambiguous in this view, for even though it seems correct to say that a small energy investment (a slight wing movement) is the cause of the action (flying toward a target), to affirm this requires a different notion of causation than that being explored in this route.

An alternative route toward characterizing the agent as the causal source of its activity could be to localize it as the center of influence in a dynamical course of events. How can the structure of influences be unpacked in a system of interactions that is complex and non-linear? It is sometimes possible to indicate the degree to which a system is affected by another at a certain point in time using statistical measures to quantify the influence that one system exerts on another or their relative independence (e.g., Lungarella, Ishiguro, Kuniyoshi, & Otsu, 2007; Seth, 2007). Typically, this involves measuring statistical correlations (in terms of predictability, mutual information, etc.) to identify statistically significant patterns and to infer causal structure from the temporal ordering of events. In this way, changes in the behavior-generating mechanisms of the system preceding environmental or relational changes in a statistically significant manner would lead us to call the system the agent of the interaction. Is this way of analyzing causal structures sufficient to describe the asymmetry between agent and environment?

It is problematic to try to disentangle the condition of asymmetry purely in terms of the structure of temporal correlations of the system/environment interaction. First, these measures rely on a series of assumptions to work (e.g., typically stationarity of the data and weak non-linearity) that might not be applicable to most of the cases. Second, and despite such correlations being clearly relevant, they might not properly describe the asymmetric relation that we sometimes find in clear cases of agency. For instance, except for what occurs in a split second interval, the physical difference between someone falling off a cliff and someone taking a dive into the ocean is probably too small to be practically captured by statistical measures of causality (unless we can count on a large sample of nearly identical instances of these events). During the fall, most of the interaction between the system (the falling body) and the environment is dominated by the environmental side (the law of gravity, the air friction, the wind), and yet one event qualifies as an action and the other does not.

It seems, therefore that we arrive at similar problems both when we attempt to characterize asymmetry in terms of energy-dependence or dynamic correlations. In some situations, these measures can indicate clear instances of agency, but in both cases, we have found situations where the proposal fails. Our solution to this situation will be to define interactional asymmetry in terms that are weaker than those of causation, but also less problematic.

We can capture the situations described above (those cases where a clear energy gathering and regulation is executed by the system in order to initiate action) and their exceptions (those cases where the agent is "surfing" the coupling with the environment) with a notion of *modulation* of the interaction. The coupling between a system and its environment is, strictly speaking, a symmetrical physical happening (but potentially with periods where one half of the coupling is statistically more dominant than the other). However, an agent is able to modulate some of the parametrical conditions and to constrain this coupling in a way that the environment (typically) does not. This condition can be expressed like this:

$$dS/dt = F_Q(S, E) \tag{1}$$

$$dE/dt = G_Q(S, E) \tag{2}$$

$$\Delta p = H_T(S) \quad p \subset Q, \tag{3}$$

where S describes the state of the agent, and E the state of the environment. Equations 1 and 2 are sym-

metrical and describe two coupled systems. The parameter Q represents a set of conditions and constraints on the coupling, including constraints internal to each system. A subset of these conditions is described at a given time by the parameter p. Equation 3 describes the asymmetrical modulation of the coupling by the agent. It applies only for an interval of time T and not for all time. For instance, before the jump, the diver is interacting with the solid ground. A sequence of muscle movements (changes in S) results in a dramatic change in the constraints that modulate the coupling with the environment (Δp) leading the system to engage in free-fall dynamics. Here we must, of course, notice that were the diver not poised at the edge of a cliff, the same sequence of muscle movement resulting in jumping forwards would produce a very different effect. This indicates that action is contextual (on Q) and temporally extended (T).

Notice that p can represent very general constraints, not just parameters that could be redescribed as variables in a way that the symmetry of coupling is restored. On the contrary, some of these constraints could be non-holonomic, that is, not describable in a manner that allows mathematical integration, the relation to the cliff edge in the case of the diver being one example.

We must also notice that the modulation of p by the agent is not necessarily a continuous occurrence, and hence such modulations are events in themselves. In addition, there are other constraints and conditions within the set Q that are not in p and so are not necessarily within reach of the agent's modulation and might be affected by other systems (either agents or not). As mentioned above, such changes are *typically* not induced by the environment, however, they can occur accidentally. An agent is a system that systematically and repeatedly modulates its structural coupling with the environment. *We therefore define interactional asymmetry as the condition describing a system as capable of engaging in some modulations of the coupling and doing so at certain times*, but not necessarily always (and, for extreme cases, just capable of halting a coupling).

There is a sense in which a proper action is realized by a coordination of the different parts of the agent not only with respect to the production of changes in p, but also with respect to the ensuing modulation being somehow aimed at a particular outcome. These are aspects that are not fully covered by the requirement of *interactional asymmetry* and so they bring us to the last condition.

2.3 Normativity

Even if we have a well-identified system being itself the active source of modulations in how it couples to the environment, we are still missing an extra ingredient in order to call it an agent. The spasms of a person suffering from Parkinson's disease are not considered to be the actions, even though the person is a well-identifiable entity and the genuine source of her interactions with the environment. Something prevents us from calling tremors or spasms actions. When considering agency we presuppose that the interaction is not random or arbitrary but makes some "sense" for the agent itself. Agents have *goals* or *norms* according to which they are acting, providing a sort of reference condition, so that the interactive modulation is carried out in relation to this condition.[2]

More knowledge than that provided by universal laws is necessary to reveal a system as an agent. It is necessary to include the contingent conditions that transform the *modulation* of the environmental coupling, into a *regulation* of the coupling, something done so as to satisfy a given norm. In other words, norms or goals cannot be deduced from universal laws alone, they show up as contingent regularities with a sense of ought-to-be in themselves: the norm *must* be followed, not doing it becomes a *failure*. Note that this is not the case for all kinds of systems. Planets cannot "fail" to follow the laws of nature. Agents, however, actively regulate their interactions and this regulation can produce failure or success according to some norm. This is what we call the *normativity condition.*

We can only make sense of norms as the result of a specific set of conditions that both enable and demand a system to distinguish between different physical outcomes of its coupling with the environment. Normativity is an essential component of agency, even if its presence can be stronger or weaker, as a degree of improvement, of increasing/decreasing adequacy, of gradual functional achievement, and so forth. This is the case independently of whether norms are linked directly or indirectly to vital requirements (the self-maintenance of the agent's biological infrastructure) or are acquired and embodied in other self-sustained forms of life (psychological, cultural, etc.). Again, it is insufficient that we, as observers, make judgments on

behalf of the agent about the "adequacy" of its behavior in relation to some of our own norms, standards, or goals (epistemic, artistic, ethical, functional, or otherwise). If we are to adopt a naturalistic approach we must be able to justify this normativity based on the very "nature" of the agent. A naturalistic but nonreductionist account of agency has to render explicit an agent's own self-defined normativity. Attempts to posit this problem and the framework under which it should be addressed have been made before (Christensen, 1999; Christensen & Bickhard, 2002; Christensen & Hooker, 2000; Di Paolo, 2005; Di Paolo, Rohde, & De Jaegher, in press; Mossio, Saborido, & Moreno, in press; Weber & Varela, 2002).

2.4 Relation Between the Three Requirements

The requirements of individuality, asymmetry, and normativity seem to characterize most of the instances in which a notion of agency is invoked in formal (e.g., legal or scientific) or informal contexts, even though more requirements must be in place to speak of more specific forms of agency. The question must be asked about the relation between these three requirements. Is it just an ad hoc list constructed so as to fit our intuitions? As it turns out, as we have already hinted, these requirements are mutually supporting ideas that altogether point to a deeper principle or organization from which they originate. But let us analyze their relationship in more detail.

The first thing to note is that the three requirements are necessary conditions for agency but none of them is sufficient on its own (nor any two of them without the third). Yet, not all of them stand in the same relationship to each other. The individuality condition appears as a precondition for the other two. Neither asymmetry nor normativity would make much sense in the lack of an individualized system to which these properties can be attributed. Even if the origin of some norms does not fully lie within the individual (e.g., social norms), it is always the individual who internalizes them, acts according to them, either succeeds or fails in doing so and, we could add, the failure or success has some effect on it (e.g., on its socially constructed identity). We have shown how once an individuality is in place there are cases in which, despite there being a clear interactional asymmetry (the case of Parkinson's tremors) there is no genuine agency without normativity. The contrary can also be the case. A system–environment coupling might be satisfying a set of norms specified by the system without the system being the agent at play. For instance a cat moving her kittens closer to her to warm them up in winter or a doctor operating on a patient. From the point of view of the system (kittens or patient) the coupling results in the accomplishment of a norm that is specified by its organization, yet the system is not the source of the modulation that is beneficial for itself. Other relations between these requirements could be explored, but these points make us think that the requirements are mutually supporting because each of them relates in non-trivial ways to the others, for instance, by covering an absence.

We have departed from an intuitive conception of agency and have argued based on common usage of terms such as actions, for the joint necessity of these three requirements. We could also add that such conditions are also sufficient for a minimal conception of agency because no additional condition seems to be required to resolve any outstanding incompatibilities with our intuitions with regards to agency in the most general sense. In other words, we find it very difficult to conceive of any system that jointly fulfills these three criteria and it is not an agent.

At this point we already have a set of criteria (albeit not generatively defined) to judge whether a system is a genuine agent or not. From a descriptive standpoint one could already make use of these conditions to evaluate whether a given system is an agent or not (see Table 1) and to test some available models. But a proper definition of agency should do something else for us: it must specify what is the generic and minimal type of system or mechanism that is capable of generating, by itself, the properties that meet these conditions. In fact all three of these requirements share in common an essential role played by the inner organization of the agent: that the system be defined by itself, that the system be active, or that it be regulating its interactions according to norms generated or sustained from within (or, in some cases, internalized by the agent). All this requires that we look inside, that we explain these features in terms of how the system is organized and organizes its interactions with the environment. As Rohde and Stewart (2008) argue, the ascription of these kinds of features on solely behavioral grounds (if possible at all) stands on a much weaker base than those grounded in the scientific study of the

Table 1

Type of system–environment interaction	*Individuality condition*: Is the system an individual?	*Interactional Asymmetry condition*: Is the system the active source of interaction?	*Normativity condition*: Is the interaction norm generated by the system?	Is the system an agent?
A gas on a container	**NO**: The gas has no identity of its own, it is an externally imposed container that limits it in space and time	**NO**	**NO**	**NO**
A cell undergoing passive osmosis	**YES**: the system produces and maintains its organization including its membrane	**NO**: passive osmosis is a physically unconstrained process it is not asymmetrically caused by the systems organization	**YES**: yes if the osmotic process is functionally beneficial in relation to the system itself (e.g., sugar concentration balanced)	**NO**
A human undergoing Parkinson tremors	**YES**: the human body produces and maintains its organization	**YES**: the system is the energetic and dynamic source of the movements	**NO**: movements are not directed or responding to any internally generated norm	**NO**
A kitten being warmed up by its mother	**YES**: the kitten is a individualized organism producing and repairing itself	**NO**: it is the mother (the environment) that is driven the coupling	**YES**: The system–environment coupling is satisfying the norm of keeping the kitten's temperature within viability boundaries	**NO**
A bacterium performing metabolic-dependent chemotaxis	**YES**: the system produces and maintains its organization including the membrane	**YES**: the system is the energetic and dynamic source of the movements	**YES**: interaction is regulated internally and directly linked to processes of self-maintenance	**YES**

underlying mechanisms involving the organization of the agent. We shall next show how minimal living organization is already capable of meeting these conditions and we will then abstract away from living organization to provide an abstract generative definition that is applicable to a wider set of contexts.

3 Living Agency

What follows (and a great part of what was previously stated) is a variation on an old theme that unifies some approaches in philosophy of mind and cognitive science by grounding cognitive capabilities in the autonomous organization of living systems. This tradition could be traced back to Aristotelian conceptions of living form and organic function, together with the Kantian interpretation of self-organization in living systems on his *Critique of Judgment.* But it was not until the rise of systems-theoretic approaches to biological organization (e.g., Bertalanffy, 1952), phenomenological approaches to philosophical biology (Jonas, 1966/2001, 1968), cybernetics (Ashby, 1960), and developmental psychology (Piaget, 1967/1971) that this tradition came closer to scientific examination and came into contact with cognitive science.

During the late 1960s and 1970s the first rigorous conceptual, mathematical, and simulation models of minimal living organization became available: Varela, Maturana & Uribe, 1974 autopoietic theory of life, Tibor Gánti's (2003a, 2003b) chemoton model, Stuart Kauffman's (1971) auto-catalytic network theory, or

Robert Rosen's (1958) M-R systems (see Rosen 1991 for an overview). The development of complexity sciences (particularly the exploration of principles of self-organization in complex networks and research in far-from-equilibrium systems in physics and chemistry) and the rise of system's biology enriched the theoretical and methodological framework of this tradition. Of particular relevance for this article is the work by Christensen and Hooker (2000) paralleled by that of Ruiz-Mirazo and Moreno (2000) where the very concept of agency is traced back to the autonomous organization of life making explicit some of the theoretical and philosophical implications (for some later developments see: Barandiaran, 2008; Di Paolo, 2005; Moreno & Etxeberria, 2005; Thompson, 2007; Van Duijn, Keijzer, & Franken, 2006). More recently, research into the minimal life forms and the origins of life (for a review see Rasmussen et al., 2008) has brought with it physically realistic simulation models and in vitro synthesized proto-cellular systems, thus providing a much more accurate and empirically grounded approach to the subject (including some relevant work for the characterization of minimal agency—see Ruiz-Mirazo & Mavelli, 2008). It is now evident that we can synthesize, simulate and analyze complete, although minimal, proto-organisms, making explicit some of the emergent (holistic and integrated) properties that are found at the root of agency and had long remained elusive to proper scientific investigation; which has favored reductionist (e.g., molecular) or abstract (e.g., mentalist) approaches.

The picture that comes out of this tradition is that the required minimal living organization is that of a far-from-thermodynamic-equilibrium system, a metabolic network of chemical reactions that produces and repairs itself, including the generation of a membrane that encapsulates the reaction network while actively regulating matter and energy exchanges with the environment. From this point of view, organisms are integrated and active systems that must continuously interact with their environment to self-generate and maintain their own dissipative organization. This minimal (or proto-cellular) living organization comes to capture the *essence* of life, for even complex multicellular organisms ultimately respond to the same logic of networked self-regeneration and self-regulation through its openness to the environment. These minimal models already provide a first empirically addressable sense of individuality and normativity without having to invoke abstract mentalistic entities such as "propositional beliefs" and "motivations" or without having to reduce the phenomenology of agency to the "selfishness" of a replicating molecule (Dawkins, 2006).

The satisfaction of the individuality condition is almost straightforward: the very organization of a living system is self-asserting, by continuously regenerating itself and its boundary a living system is demarcating itself from its surroundings as a unified and integrated system. In doing so it also carves an environment out of an undifferentiated surrounding: the organization of the system (the way in which components processes are nested with each other building up a whole) determines which environmental features are "relevant" to it, that is, which chemical components in the environment can affect it or are needed for its continued existence. In this way, the environment is not just what lies outside the system as demarcated from the observer's point of view but is specified by the system through the set of boundary conditions that affect it. The system as a whole is irreducible to the sum of its disjoints parts, as few reactions of a protocellular system would occur in the absence of the continued support provided by the networked reactions as an organization (that produce catalysts and molecular components at the appropriate rate—sustaining reactions away from their thermodynamic tendency) and the presence of a membrane that acts both as a container and regulator of these reactions.

In turn, this is where living individuality naturally leads to normativity: component reactions *must* occur in a certain manner in order for the very system to keep going, environmental conditions are *good* or *bad* for the continuation of the system, the system can *fail* to regain stability after a perturbation, and so forth. This normative dimension is not arbitrarily imposed from the outside by a designer or external agent that monitors the functioning of the system and judges according to her interests. It is the very organization of the system that defines a set of constraints and boundary conditions under which it can survive (Barandiaran, 2007, 2008; Christensen & Bickhard, 2002; Mossio et al., in press). In this sense, living systems are subject to a permanent *precariousness* (Di Paolo, 2009) that is compensated by its active organization. This precariousness implies that whatever the organism is doing (i.e., whatever its factual functioning is) there is something that it *ought* to do; not for an external observer but for itself, for the continuation of its very existence.

In Jonas' words: "[for metabolism] 'To be' is its intrinsic goal. Teleology comes in where the continuous identity of being is not assured by mere inertial persistence of a substance, but is continually executed by something done, and by something which has to be done in order to stay on at all: it is a matter of to be or not to be whether what is to be done *is* done." (Jonas, 1968, p. 243). This type of organization we call *autonomous* (following Varela, 1979—for later developments see the contribution in Barandiaran & Ruiz-Mirazo, 2008), since it captures both the emergence of a self (*autos*) and that of norms (*nomos*).

The permanent need for external matter and energy, and the fragility of living systems, sooner or later leads to interactional asymmetry: any organism must actively seek energy gradients and regulate its relation with the environment in order to compensate or avoid potentially destructive perturbations. So, over the most minimal metabolic network endowed with a membrane, even very simple life forms possess adaptive mechanisms that operate detecting and regulating internal and interactive processes. Ruiz-Mirazo and Moreno argue that the ion pumping mechanism against chemical gradients provides one of the simpler examples: internal concentrations are regulated by modulating membrane permeability according to self-maintenance conditions (Ruiz-Mirazo & Mavelli, 2008; Ruiz-Mirazo & Moreno, 2000). But paradigmatically, it is chemotactic behavior moving up metabolic substrate gradients, or moving down poisonous reactant ones, that brings simple life forms closer to our intuitive notion of agency: the system is coupled to the environment through a specialized (yet metabolically modulated) sensorimotor subsystem capable of engaging in interactive cycles whose modulation (in terms of changing the frequency of the direction of rotation of its flagella) becomes essential for the metabolic continuation of the bacterium.

Minimal life forms already come to satisfy the necessary and sufficient conditions for agency. This does not imply, however, that living organization is necessary for agency, nor that all forms of agency need to trace their normative or individuality conditions back to living organization. What minimal life provides is a clear and precise illustration of how individuality, normativity, and interactional asymmetry conditions emerge from a naturalized framework that can be fully operationalized and even synthesized. What is essential for agency is that, in a manner isomorphic or analogous to that of metabolism, interactive processes can be traced back to a form of organization that displays similar properties.

4 Defining Agency

We have highlighted a number of *conditions* for agency and we have illustrated how living organization might come to satisfy them as one minimal (but not necessarily unique) example. We have based our analysis on the common usage of the terminology related to agency. This has helped us avoid providing a definition that could be at odds with our intuitions about agency. We are now in place to provide a *generative definition* that is not just a restatement of the conditions, but also the description of an organization capable of generating and satisfying them. In short, an agent *is an autonomous organization capable of adaptively regulating its coupling with the environment according to the norms established by its own viability conditions.* (Figure 1 depicts the relation between the different elements that make up the definition.) A more detailed and complete definition goes as follows:

A system *S* is an agent for a particular coupling *C* with an environment *E* iff:

1. *S* is an *open autonomous* system in an environment *E*, meaning that:
   - 1.1. among a set of processes a system *S* can be distinguished as a network of interdependent processes whereby every process belonging to the network depends on at least another process of the network and enables at least another one so that isolated from the network any component process would tend to run down or extinguish;
   - 1.2. the set of processes (not belonging to *S*) that can affect *S* and are affected by *S* defines *S*'s environment (*E*); and
   - 1.3. *S* depends on certain conditions (specified by *S*) that in turn depend on *E*
2. *S modulates* the coupling *C* in an *adaptive* manner:
   - 2.1. where *modulation* indicates an alteration (dependent on *S*) in the set of constraints that determine the coupling between *S* and *E*;
   - 2.2. *adaptive* means that the change in the coupling *C* contributes to the maintenance of some of the processes that constitute *S*.

The above definition has two main virtues: (a) it is not circular (the terms used in the definition do not presuppose the notion of agency) and (b) it is generative (in the sense that the three requirements previously stated for agency follow from the definition). We shall first show how the individuality, normativity, and interactional asymmetry conditions are satisfied and generated by the definition and then we will try to clarify some of the terms included in the definition.

The autonomous organization (statement 1) provides a concrete sense of individuality and normativity. Statement 1.1 provides an objective criterion to *individualize* a system among a set of processes: those networked interdependently constitute an individual. Although the network also depends on environmental processes, these are not part of the system because they do not, in turn, depend on the system (although they might be affected by it). Since the component processes cannot be sustained by themselves but only through the network of dependencies to other processes it follows that the system is also *self-sustaining* in the sense that the organization not only defines the system but it is also thanks to it that the system endures in time. *Normativity* emerges from how the constitutive processes and the dependencies between the system and the environment (described in 1.1 and 1.3) affect self-maintenance. Specific norms relate to the different ways in which a change in the system's processes or in the environment can lead the system to lose its organization as a self-maintaining network.

Since the system modulates its coupling with the environment in relation to conditions of self-maintenance (statement 2), *interactional asymmetry* is guaranteed. There is a specific sense in which the system can be said to be the source of the actions, for not only is it modulating the coupling but is doing so in relation to the norms; that is, it is the organization of the system (from which norms emerge) that is determining the modulation of the coupling. When we consider that the system depends on certain conditions (1.2 and 1.3) and that it modulates its coupling in relation to them (2.2) a deep sense of agency comes to the surface: the system is not only acting (2.1) but it is through its actions that it maintains and individuates itself.[3]

It is the deep circularity and entanglement between networked processes, the self-maintaining conditions they generate, and the interactions that the system establishes with the environment that makes agents so challenging to model and understand. To make precise sense of such entanglement requires to make explicit some of the terms involved in the definition. For instance, the concepts of "maintenance," "enabling," or "dependence" play a key role. Such terms were chosen because they have a considerable degree

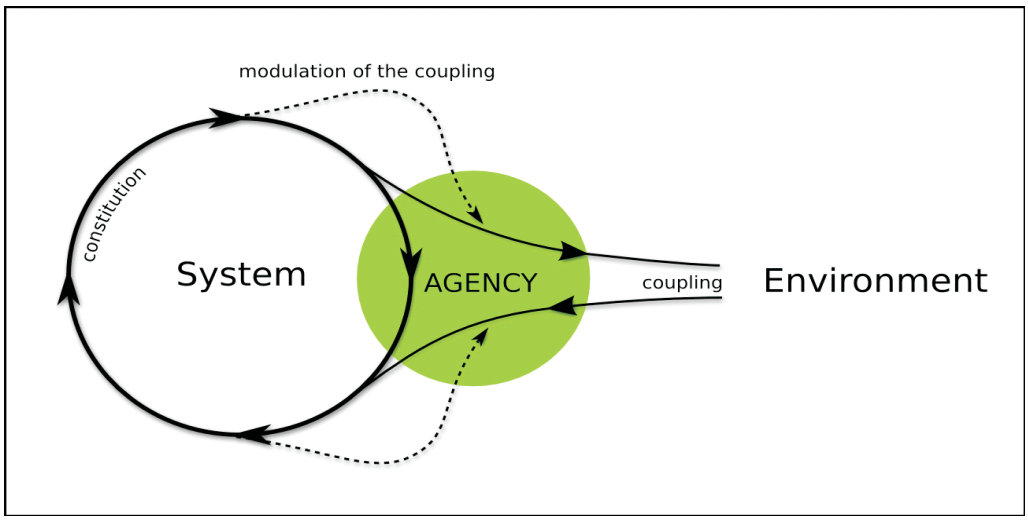

Figure 1 The figure illustrates the definition of agency: the system is constituted by a self-sustained network of processes (pictured as a circle, left) coupled to the environment; the systems exerts regulatory constraints over its coupling giving rise to agency. (Copyright © 2009 Xabier E. Barandiaran under a Creative Commons Attribution Share-Alike license, freedom is granted to copy, modify and distribute provided that this notice is preserved.)

of generality and may take different forms in specific cases. Sometimes such terms can be interpreted in terms of stability (or in relation to a certain dynamic order). For instance, the *maintenance* of a process might refer to its stability around a given region of its state space (e.g., an attractor); in turn, a process *X* would *depend* on another process *Y* if *Y* is a parametric condition of *X*'s stability (and, conversely, *Y* would *enable X*'s stability). In other cases a process is providing the material or energetic requirements for another process to be sustained; or it might constrain its degrees of freedom in order to operate in a certain manner. Other cases might involve more sophisticated forms of dependency.

Another central concept of the definition is that of component *processes*. Unlike other authors (particularly Ruiz-Mirazo & Moreno, 2000, and Christensen & Hooker, 2000) we do not restrict or reduce autonomy to the domain of metabolism or biological organization. Our definition of autonomy (much in the line of Varela, 1979) can be applied to other domains. For instance, networked interdependent processes can be chemical reactions, molecular structures, physiological structures (such as tissues or organs), neurodynamic patterns at the large scale, sensorimotor loops, social habits, and so forth. This way, agency does not have to be subordinated to biological/metabolic organization but can appear at different scales responding to a variety of autonomous processes. The possibility is also open for different forms of autonomous organizations to overlap in their material substrates.

What remains central to our definition is that for any agentive engagement of a system with its environment its identity must be jeopardized at the proper level and that the interaction must involve a process of compensation for deviations from a norm that is generated from within (both the norm and the compensation). It is in this sense that the interaction becomes meaningful for the agent, that the agent makes *sense* of a situation: actions are guided by the need to compensate the threatening deviation from a norm and environmental processes are integrated into the interaction as *relevant* for the achievement of such compensation. We call this process *sense-making* (Di Paolo, 2005; Thompson, 2004) for what would otherwise be a mere event or occurrence becomes *valued.* The threat must not be interpreted exclusively in terms of a direct challenge to the continuation of the agent. It can take the form of a tension or imbalance that, without directly challenging the identity of the system, still provokes an involvement of the whole system in its attempt to counteract the imbalance with the effect that more direct threats are consequently averted.

The definition we provide is minimal, it is just meant to generate phenomena that fulfill necessary and sufficient conditions for agency (at whatever level of organization that is considered). Notice for instance that living systems in conditions that require interactive regulation fall under this definition, consequently life is sufficient for agency (but not necessary because, as we mentioned, the possibility is open for non-metabolizing forms of autonomy and agency to appear). While capturing those features that are essential to minimal forms of agency, the definition remains open to further conditions, interdependencies, hierarchies of modulation, forms of coupling, and so forth. that might account for more complex types of agency. Similarly, we should not expect natural agents to operate at a single level of organization. Most will involve different scales of autonomy (metabolic, immune, neurodynamic, social, etc.) forming nested hierarchies of adaptive regulation (such as metabolic monitoring mechanisms modulating behavioral responses or neurodynamically induced psychosomatic disorders in the immune system). But leaving aside the sophisticated cases that involve different scales of autonomy it is fundamentally through the spatial and temporal dimensions that agency expands in complexity.

5 The Spatio-temporal Dimensions of Agency

Agency is inherently a temporally and spatially extended process. When we say so, we mean not only that the processes described have an essential temporal or spatial extension in the eye of the observer, but also that an agent's own perspective has temporal and spatial structure and that this depends on its form of agency. In this section the analysis is less rigorous than in the preceding sections, as our purpose is to address, as an example, not as an exhaustive treatment, some of the consequences of the previous minimal definition and how it can be made more complex along the fundamental dimensions of space and time. Most of these issues remain open to further work.

5.1 Why Space–Time Matters

It is evident, even trivial, that time is an essential aspect of agency. Following our definition we first notice that it is through interconnected *processes* that an autonomous organization is constituted (a frozen snapshot of the system is nothing but a picture of a dead organization). The time span of the interdependencies between such processes and their precariousness (their extinction outside the organization that sustains them) is also crucial to understand the self-maintenance of the system and its margin to compensate decay and perturbations. In addition, different rhythms, temporal scales, and phenomena of synchronization and co-variation might be found at the core of constitutive processes (Buzsáki, 2006). Second, we also notice that the *adaptive modulation* of a coupling makes agency unfold temporally: in order for a system to regulate itself there must be some buffering or distance between the immediate perturbation and the possibility of compensating for it. The very notion of modulation is inherently temporal, it directly implies a change over time and often involves more sophisticated temporal structures such as a change on a certain rate (e.g., catalysts modulating a reaction rate) or the entrainment of certain rhythms by other ones (e.g., frequency and amplitude modulation). In addition, an action "is a structured event, with clearly defined phases of onset (the sensing of a negative tendency), acceleration (the activation of the adaptive mechanism), consummation (the overturning of the negative tendency) and cadence (the de-activation of the adaptive response)" (Di Paolo, 2005, p. 442). The temporal structure of behavior can appear as a nested hierarchy of modulations that in turn becomes crucial if we are to differentiate between types of couplings: actions, tasks, training, learning, development, and so forth.

There is also a sense in which spatiality turns out to be relevant for many forms of agency (certainly for living systems): that is, the spatial or topological properties of the processes that constitute the autonomous organization of the system and also its coupling. On the one hand constitutive processes (and interdependent relationships between them) might crucially rely on spatial aspects; for instance the formation of spatially structured patterns in self-organized processes such as convection flows (Hanczyc, Toyota, Ikegami, Packard, & Sugawara, 2007).

But more important, perhaps, is the spatial situatedness of agents. The need to understand agents as brain–body–environment coupled dynamical systems has been repeatedly stressed (Beer, 1995, 2003; Chiel & Beer, 1997). But a motile sensorimotor coupling is a special case of coupling, qualitatively different from the paradigmatic example of the coupling of a Watt's Governor with a water flow. A sensorimotor coupling is, primarily, a coupling between a geometrical space and a dynamical system. This implies, first of all, that behavior cannot be taken to be exclusively the result of extracting statistical properties or patterns from a string of predefined sensory inputs and the production of an adequate response output. Situatedness provides much more complex and flexible possibilities for action. For example, a non-situated agent whose control architecture is reactive (i.e., whose output is determined by the instantaneous input by a non-modifiable internal structure) cannot solve a non-Markovian task, that is, cannot successfully classify an environmental condition if its detection requires the extraction of a sequential (timely) order, when the condition of the environment cannot be reduced to an instantaneous sensory value. In contrast, a system with a reactive controller that is situated in a spatial environment can transform non-Markovian tasks into Markovian tasks by exploiting geometric properties of the agent–environment coupling (e.g., Izquierdo-Torres & Di Paolo, 2005). Yet, what is particularly relevant for agency is the establishment of a *perspective*, that is, the constitution of the agent as a reference point in space and time, for the agent itself and not for the observer (who can chose any reference point for its convenience and accommodate the observations accordingly). With this *perspective* a proper sense of environment for the system, or world, emerges, not just a surface in which external processes impinge, but a spatially and temporarily structured domain of interactions.

5.2 Space–Time from the Agent's Perspective

Poincaré (1895) has argued that the Euclidean geometrical properties of an agent's world are to the result of its sensorimotor situatedness in a spatial environment and to its capacity to enact invariant properties (such as continuity of space, dimensionality, or homogeneity) through sensorimotor structuring of its experience (such as active visual tracking, reversibility of perceptions, and invariance of shape upon movement around an object). Even when he was not directly concerned

with the nature of agency Poincaré conceptualized spatial properties as arising from the possibilities and regularities of bodily actions. Motility in a spatial environment equips an agent with the possibility of coping recurrently with the perturbations it encounters and to span them onto a domain of interactions and flexible sensorimotor correlations. With such a way of recurrent modulation, an agent has the possibility of restoring situations at will, exploiting the structural invariants of the sensorimotor coupling with the environment that it thus creates. Therefore, the challenge is to reconstruct the spatio-temporal dimensions of the environment of the agent not from the point of view of the observer scientist or the modeler, but from the frame of reference of the agent itself. We may know as external observers that an agent and its environment have certain spatial properties (such as a number of coordinates, objects, gradients, etc.). We may also make these properties part of our models, for ease of interpretation and familiarity. However, on many occasions, for an agent, the environment could be reduced to a string of intensities with no spatial or temporal patterns—the spatial and historical structure is in the eye of the observer.[4]

The spatial and temporal organization of behavior is the result of the lifetime development of spatio-temporality (as, e.g., described by Piaget, 1946/1969; Piaget & Inhelder, 1948/1956) and the elaboration of spatio-temporal behavior in evolutionary history. Such an elaboration of space and time in an agent's behavioral domain (from its own perspective) coincides with higher forms of agency with more complex mechanistic organization. More complex life forms show a build-up of both quantitative growth and qualitative transitions of spatio-temporality and agency: rudimentary forms of memory provide a sequentially ordered space without a metric—for instance, Cartwright and Collett (1983) have proposed that the mechanism used by honeybees to return to the hive after foraging comes down to the restoration of an ordered sequence of retinal snapshots. Some insects, mammals, and birds clearly exploit not just the orderly, but also the metric properties of their couplings with the environment. Their rich sensorimotor inventory, afforded by the nervous system's fast and flexible way of linking sensors and actuators, allows them to further increase the degree of mediacy between the surface effect of the stimulus and its meaning for the system by adding another layer of abstraction to its perspective on the spatio-temporality of its coupling with the environment. This transition in spatio-temporality coincides with a transition in agency.

Human spatio-temporality, even though it is still not well understood in many ways, adds yet another layer of abstraction to the perspective on the spatio-temporality of the environment. This additional layer is the symbolic abstraction of space and time. For instance, we tend to think of space as an abstract, unitary three-dimensional box. The reality of our embodied behavior shows, by contrast, that our interactions with the world in the vertical dimension are strongly influenced by the vestibular sense (because of gravity), which makes them very different from our interaction with the world in the horizontal plane (e.g., Gibson, 1952). Similarly, we make an explicit spatial analogy of time as an arrow in thinking and language (cf., Lakoff & Johnson, 2003; Rohde, 2008). Such symbolic spatio-temporality, that lumps together a diverse set of sensorimotor couplings with the world pushes the stimulus and its meaning even further apart. We tend to think of this Euclidean perspective on space and time as the "real" one—because it is the one that most makes sense in our behavioral domain, afforded by our body, its coupling with the environment and cultural practices. One of the most difficult and challenging tasks to advance the understanding and modeling of agency is to reconstruct this process of spatio-temporal organization of agency without directly mapping or projecting our human preconceptions into the agents we study.

6 Conclusions

We have attempted to describe a phenomenon that is both commonplace and practical, especially in research areas such as cognitive science and robotics, and yet very hard to define. The generative definition that we have provided answers to the necessary and sufficient conditions that we have identified for agency. Yet there are many aspects of the definition (and the requirements) that need further work. For instance the characterization of interactional asymmetry in terms of modulation is still incomplete. It would be valuable to investigate how energetic and informational/correlational measures of asymmetry relate to modulatory capacity and to explore whether such measurements can complement each other. In addition, a potential

problem of our formalization of modulations is that what counts as a parametric modulation and what as a coupling between variables is often a matter of choice for the modeler. Therefore, a principled way to account for interactional asymmetry would be highly desirable. An asymmetry between the complexity of behavior generating mechanisms and that of the coupling could be an alternative, but most probably to move a step further requires connecting the issue of interactional asymmetry with that of normative regulation.

Other questions have to do with the relationship of co-dependence between system and environment. Although a first approximation of the problem required distinguishing the system from its environment, agency (especially when considering recurrent sensorimotor situatedness) leads to a deep entanglement of an agent with its environment. Yet, despite its "being-in-the-world" an agent does *selectively* couple with environmental features asymmetrically integrating them on its behavioral organization. A number of questions follow: How does niche construction (for example) relate to agency? Should those environmental features that recurrently depend on the agent be considered as part of the agent? What is the status of tools as mediators between agents and environments?

Finally, the relationship between the emergence of norms and adaptive regulation also requires a more careful analysis. How, exactly can a system be "sensitive" or "responsive" to its own norms when these emerge from holistic dependencies? What is the status of living behavior that is produced by mechanisms that operated independently of the rest of the organisms and yet do so in an adaptive manner (e.g., a reflex)? This issue also connects with a requirement that is somehow implicit in the definition: that the agent be involved, as a whole, in the modulation of its coupling and not driven by subsystems operating as central controllers for the rest of the system.

To address these issues might require the use of specific models (natural or artificial) where questions can be more precisely formulated, alternatives tested on the model, and conceptual issues clarified in a workable arena (capable of rendering explicitly the consequences of complex mechanisms in operation). Similarly, applying the definition to an existing model (illustrating aspects of the definition with specific processes in the model) would also permit the testing of potential measurements of the notions of norms (e.g., homeodynamic stability), individuality (e.g., with measures of closure based on hyperset or category theory), and interactional asymmetry (e.g., in terms of emergent constraints or some measure of complexity asymmetry). Also, to attempt a formalization of the definition (or of some aspects of it) would be a parallel endeavor to that of developing appropriate measures. Ultimately, the ideal definition should be able to confront the following task: given a natural (physicochemical or biological) system or a mathematical/simulation model (where variables and parameters do not bear any information about what they are meant to represent) we should be capable of using the definition almost automatically (i.e., without requiring further interpretation) to distinguish between the system and the environment, to define the norms that the agent must satisfy and to determine whether the system is operating as an agent or not.

Despite the fact that our definition is, admittedly, not yet complete there are concrete and practical consequences that can be extracted for the study of adaptive behavior: (a) mere sensorimotor coupling on its own is too weak a condition for agency, modulation of interactions need also be considered; (b) systems that *only* satisfy constraints or norms imposed from outside (e.g., optimization according to an externally fixed function) should not be treated as models of agency; and (c) the identity of an agent cannot be divorced from its behavior, therefore, some kind of feedback between the agent's behavior and the self-maintenance of its organization should be included in our models (i.e., the agent must "benefit" or "suffer" the consequences of its action in a manner that is relevant for its continued activity). Finally, it must be stressed that models of agency can explore different aspects of our definition without the system fully satisfying the three requirements.

Stressing the importance of space and time also has relevant implications for modeling and developing theoretical intuitions about agents: while some of the temporal aspects have begun to be investigated more systematically with dynamical systems approaches, the complexities brought by spatial embeddedness are still not fully captured by traditional modeling frameworks.

Through its emphasis on the dynamic-organizational requirements that make a system an agent our definition of agency provides a step toward a workable scientific concept that can be used to regulate empiri-

cal and modeling work. It is not, however, expected to fit all potential legitimate uses of the term, hence we prefer to see it as a definition of *minimal* agency. It is open to further theoretical development and revision in the light of modeling and empirical advances. For empirical research, our definition indicates that it will be important to understand how organizational autonomy links to motion and spatio-temporal behavior, but also how these domains are distinguished at the level of the generative mechanisms, and what that implies for the kind of patterns an agent can regulate, its possibilities for action and perception, and what they mean for an agent's spatio-temporally organized sense-making. Our definition also points out the routes towards the synthesis of such agents, the kind of mechanical-dynamical organization to aspire. We expect progress to come from minimal models that explore the dynamic relationship between constitutive aspects of the agent (those involved in the maintenance of an autonomous organization) and interactive aspects (coupling with the environment and the modulation of that coupling). In return, these models and their analysis will indicate a need for improvement and enable us to progressively distill an increasingly workable and formalized definition.

As a further observation, we note that existing models of autonomous agents either focus on constitutional aspects alone or they focus exclusively on interactive aspects independently of how they relate to constitution, that is, simple sensorimotor loops. Daley et al. (2002), Fernando and Rowe's (2008) models focus on the chemical requirements to build an individual but do not address interactive aspects. Concerning the other end, typified by agent modeling in autonomous robotics and minimal cognition, even those researchers concerned with the connection between viability constraints and behavior, have not focused on the question of how sensorimotor loops relate to autonomous constitution (though some of the relevant questions can indeed be approached with these methods, see, e.g., Di Paolo & Iizuka, 2008; Iizuka & Di Paolo, 2007).

There are, however, an increasing number of models that explore the intermediate space and issues of modulation. Some of them draw their models departing from metabolic organization: they include models by Ruiz-Mirazo and Mavelli (2008), Ikegami & Suzuki, (2008), and Hanczyc et al. (2007). Other recent models explore the issues of decoupling between sensorimotor mechanisms and metabolic dynamics that our definition points to (e.g., Egbert & Di Paolo, this issue) and can be situated even closer to the idea of agency expounded here. Certainly, the grounding of individuality and normativity conditions in biological organization and minimal models of metabolism has attracted most attention. But it has also distracted attention away from an almost unexplored avenue of research: the possibility for the emergence of a new level of autonomy in the domain of behavior and neurodynamics. The adaptive regulation of behavior needs not be exclusively subordinated to the viability constraints imposed from biological "survival conditions." Instead, it can be equally governed by the need to maintain neurodynamic and behavioral organization in terms of self-maintenance of habits, coherence of behavior, psycho-dynamic stability, and so forth (Barandiaran & Di Paolo, 2008; Barandiaran & Moreno, 2006; Di Paolo, 2003;). It is here where robotic models have an opportunity to address a strong conception of agency, but further work is required to put the present definition to practical use in this direction.

There are numerous issues around the concept of agency that we have not even addressed yet: notably, the multiplicity of agency, collective agency, social, and cultural norms, or specific forms and orders of agency (intentional, reflexive, socio-linguistic, etc.). But by sketching out the most minimal and fundamental generative definition for agency, we want to provide the groundwork for future advances that allow us to tackle these problems hands on. We expect that the mentioned and future models, and a revision and extension of empirical evidence will help us eliminate potential problems in our definition and consolidate the understanding of agency.

Being specific about the requirements for agency has told us a lot about how much is still needed for the development of artificial forms of agency but at the same time it gives an indication of the different goals that must be achieved along the way. It also gives us an idea of the difficulties and complexity of the problems that still need to be resolved. This situation contrasts with the much more vague (but widespread) targets that are usually presented as goals for AI and robotics research such as asking a system to be rational, to have real emotions or to be conscious. Such targets are unfortunately very common and tend to come together with a poor mapping of the intermediate stages required to define a fruitful research program, Further-

more, they all presuppose, rather uncritically, that there exists an "unproblematic" substrate for such properties: an agent.

Notes

1 This remark applies to agents once they are in full enjoyment of their agential character. But it does not preclude the possibility that the ontogeny and evolution of different forms of agency is not itself highly dependent on links to a community of other agents and environmental factors. A self-defined identity does not happen in a vacuum and is inevitably tied to a web of necessary relations to develop and survive. The full sense of a system self-defining its own identity will be clarified in sections 3 and 4.

2 We shall use the terms *norm* and *goal* interchangeably. Despite the notion of norm being generally applied to a procedure or a limit condition that must be respected whereas that of goal refers to specific reference states (get to position X, grasp object Y, attain result Z), for minimal cases both terms might be treated equivalently since both capture a *necessary* or *desired* condition that a process must achieve. Explicit distinctions between norms, rules, goals, intentions, desires, plans, and so forth. would demand reference to more elaborate forms of agency that remain out of the scope of this article.

3 There is an additional feature of agency that is implicit in the definition but still remains difficult to fully disclose. The adaptive modulation of constraints of the coupling (which may also be global constraints on the operations of the parts of the system) must be emergent (meaning that it cannot be attributed to a sub-system of the agent in charge of regulation, a kind of central controller). This seems to be a question of logical necessity. Subsystems that regulate according to a fixed norm independent of the agent's current state and dynamical organization act as external constraints on the system, they are outside its sphere of influence. Therefore, only the emergent modulations of constraints that concern the system as a whole can be sensitive to the norms defined by the system.

4 Consider the case of *E. coli* performing chemotaxis on a chemical gradient. From the point of view of the bacterium the environment appears simply as the succession of the changes of the intensity of a variable (concentration of the attractant) with no metric structure. These changes covary with the activity of flagellar rotation. Because of the internal dynamics of the sensorimotor system this variable will tend to have, statistically, a higher value over time. The fact that the bacterium is situated in a spatial gradient is absolutely irrelevant from the point of view of the bacterium. In other words, the bacterium does not have access to a notion of its own displacements in space. Thus, the spatial properties often obvious to us, observers, of the behavior of an agent need not accessible to the agent itself.

Acknowledgments

We are grateful to Alvaro Moreno and two of the anonymous reviewers for their helpful comments. Xabier Barandiaran and Ezequiel Di Paolo have received funding for collaboration from grant HUM2005-02449/FISO of the Spanish Ministry of Education. Xabier Barandiaran currently holds a post-doctoral fellowship from Ministerio de Educación y Ciencia, Programa Nacional de Movilidad de Recursos Humanos del Plan nacional de I-D+I 2008-2011. Xabier Barandiaran also wants to acknowledge the economic support of the *euCognition* network for his position at the Autonomous Systems Lab (Universidad Politécnica de Madrid, Spain) and to the Konrad Lorenz Institute for the study of Evolution and Cognition where he developed part of the work contained on this article. Marieke Rohde would like to thank the JSPS (short term postdoctoral fellowship) and the HFSP for funding the research presented.

References

Ashby, W. R. (1960, second edition). *Design for a Brain. The origin of adaptive behaviour*. Chapman and Hall, London, UK.

Barandiaran, X. (2007) Mental life: conceptual models and synthetic methodologies for a post-cognitivist psychology. In B. Wallace, A. Ross, T. Anderson, & J. Davies. (Eds.), The Mind, the Body and the World: Psychology After Cognitivism? (pp. 49–90), Exeter, Imprint Academic.

Barandiaran, X. (2008). *Mental life. A naturalized approach to the autonomy of cognitive agents*. Unpublished doctoral dissertation, University of the Basque Country, Spain. http://barandiaran.net/phdthesis/

Barandiaran, X., & Di Paolo, E. A. (2008). Artificial mental life (Abstract). In S. Bullock, J. Noble, R. A. Watson, & M. A. Bedau (Eds.), *Proceedings of the 11th International Conference on Artificial Life, Alife XI*. Cambridge, MA: MIT Press.

Barandiaran, X., & Moreno, A. (2006). On what makes certain dynamical systems cognitive. *Adaptive Behavior*, *14*, 171–185.

Barandiaran, X., & Ruiz-Mirazo, K. (Eds.) (2008). Special issue on Modelling Autonomy. *BioSystems*, *91*(2).

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, *72*, 173–215.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, *11*, 209–243.

Bertalanffy, L. (1952). *Problems of life: An evaluation of modern biological thought.* London, New York: J. Wiley & Sons.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.

Buzsáki, G. (2006). *Rhythms of the brain*. Oxford: Oxford University Press.

Cartwright, B. A., & Collett, T. S. (1983). Landmark learning in bees. Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology, 151(4), 521–543. doi: 10.1007/BF00605469.

Chiel, H. J., & Beer, R. D. (1997). The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neurosciences*, *20*, 553–557.

Christensen, W. D. (1999). *An interactivist-constructivist approach to adaptive intelligence and agency*. Unpublished doctoral dissertation, University of Newcastle, Australia.

Christensen, W. D., & Bickhard, M. H. (2002). The process dynamics of normative function. *Monist*, *85*(1), 3–28.

Christensen, W. D., & Hooker, C. (2000). Autonomy and the emergence of intelligence: Organised interactive construction. *Communication and Cognition – Artificial Intelligence*, *17*(3–4), 133–157.

Daley, A. J., Girvin, A., Kauffman, S. A., Wills, P. R., & Yamins, D. (2002). Simulation of chemical autonomous agents. *Z. Phys. Chem.*, *216*, 41–49.

Dawkins, R. (2006). *The selfish gene*. 30th Anniversary edition. Oxford: Oxford University Press.

Di Paolo, E. A. (2003). Organismically inspired robotics: homeostatic adaptation and natural teleology beyond the closed sensorimotor loop. In K. Murase & T. Asakura (Eds.), *Dynamical systems approach to embodiment and sociality* (pp. 19–42). Adelaide: Advanced Knowledge International.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, *4*(4), 429–452.

Di Paolo, E. A., and Iizuka, H., (2008). How (not) to model autonomous behaviour, BioSystems 91: 409–423.

Di Paolo, E. A. (2009). Extended life. *Topoi*, *28*, 9–21.

Di Paolo, E., Rohde, M., & De Jaegher, H. (in press). Horizons for the enactive mind: Values, social interaction, and play. In J. Stewart, O. Gapenne, & E. A. Di Paolo (Eds.), *Enaction: Towards a new paradigm for cognitive science*. Cambridge, MA: MIT Press.

Fernando, C. T., & Rowe, J. (2008). The origin of autonomous agents by natural selection. *Biosystems*, *91*(2), 355–373.

Franklin, S., & Graesser, A. (1996). Is it an agent, or just a program? A taxonomy for autonomous agents. In *Proceedings of the Workshop on Intelligent Agents III, Agent Theories, Architectures, and Languages*, Lecture notes in computer science (Vol. 1193, pp. 21–35). Berlin: Springer.

Gánti, T. (2003a) The Principles of Life. Oxford University Press.

Gánti, T. (2003b) Chemoton Theory: Theory of Living Systems, New York: Kluwer Academic/Plenum Publishers.

Gibson, J. J. (1952). The relation between visual and postural determinants of the phenomenal vertical. *Psychological Review*, *59*, 370–375.

Hanczyc, M., Toyota, T., Ikegami, T., Packard, N., & Sugawara, T. (2007). Chemistry at the oil-water interface: Self-propelled oil droplets. *Journal of the American Chemical Society*, *129*(30), 9386–9391.

Hendriks-Jansen, H. (1996). *Catching ourselves in the act: situated activity, interactive emergence, evolution, and human thought*. Cambridge, MA: MIT Press.

Iizuka, H., and Di Paolo, E. A. (2007). Toward Spinozist robotics: Exploring the minimal dynamics of behavioural preference, Adaptive Behavior, 15:359–376.

Ikegami, T., & Suzuki, K. (2008). From homeostatic to homeodynamic self. *BioSystems*, *91*(2), 388–400.

Izquierdo-Torres, E., and Di Paolo, E. A. (2005). Is an embodied system ever purely reactive? In M. Capcarrere, A. A. Freitas, P. J. Bentley, C. G. Johnson, J. Timmis (Eds) ECAL 2005, LNAI 3630 Berlin: Springer Verlag, pp. 252–261.

Jonas, H. (1968). Biological foundations of individuality. *International Philosophical Quarterly*, *8*, 231–251.

Jonas, H. (2001). *The phenomenon of life: Toward a philosophical biology*. Evanston, IL: Northwestern University Press. (Original work published in 1966 in English)

Kauffman, S. (1971). Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. *Journal of Cybernetics*, *1*, 71–96.

Kauffman, S. (2000). *Investigations*. Oxford: Oxford University Press.

Lakoff, G., & Johnson, M. (2003). *Metaphors we live by*. Chicago: University of Chicago Press.

Lungarella, M., Ishiguro, K., Kuniyoshi, Y., & Otsu, N. (2007). Methods for quantifying the causal structure of bivariate time series. *International Journal of Bifurcation and Chaos*, *17*(3), 903–921.

Maes, P. (1994). Modelling adaptive autonomous systems. *Artificial Life*, *1*, 135–162.

Maturana, H. & Varela, F. J. (1980). *Autopoiesis and cognition*. Dordrecht, Holland: Reidel. (Original work published in 1972 in English)

Moreno, A., & Etxeberria, A. (2005). Agency in natural and artificial systems. *Artificial Life*, *11*(1–2), 161–176.

Mossio, M., Saborido, C., & Moreno, A. (in press). An organizational account of biological functions. *The British Journal for the Philosophy of Science*.

Pfeifer, R., & Scheier, C. (1999). *Understanding intelligence.* Cambridge, MA: MIT Press.

Piaget, J. (1969). *The child's conception of time*. (A. J. Pomerans, Trans.). London: Routledge & Kegan Paul. (French original work published in 1946)

Piaget, J. (1971). *Biology and knowledge: An essay on the relations between organic regulations and cognitive processes*. Edinburgh: Edinburgh University Press. (original work published in 1967, translated by Beatrix Walsh)

Piaget, J., & Inhelder, B. (1956). *The child's conception of space*. New York: Norton. (Original work published in 1948 Translated from the French by F. J. Langdon, J. L. Lunzer)

Poincaré, H. (1895). L'espace et la géométrie. *Revue the métaphysique et de morale*, *3*, 631–646.

Rasmussen, S., Bedau, M. A., Chen, L., Deamer, D., Krakauer, D. C., Packard, N. H., & Stadler, P. F. (2008) *Protocells. Bridging nonliving and living matter*. Cambridge, MA: MIT Press.

Rohde, M. (2008). *Evolutionary robotics simulation models in the study of human behaviour and cognition*. Unpublished doctoral dissertation, Department of Informatics, University of Sussex. http://mariekerohde.com/contents/Rohde2008.pdf

Rohde, M., & Stewart, J. (2008). Ascriptional and "genuine" autonomy. *BioSystems*, *91*(2), 424–433.

Rosen, R. (1958). A relational theory of biological systems. *Bulletin of Mathematical Biophysics*, *20*, 245–341.

Rosen, R. (1991). *Life itself. A comprehensive enquiry into the nature, origin and fabrication of life*. New York: Columbia University Press.

Ruiz-Mirazo, K., & Mavelli, F. (2008). On the way towards "basic autonomous agents": Stochastic simulations of minimal lipid–peptide cells. *BioSystems*, *91*(2), 374–387.

Ruiz-Mirazo, K., & Moreno, A. (2000). Searching for the roots of autonomy: The natural and artificial paradigms revisited. *Communication and Cognition-Artificial Intelligence*, *17*(3–4), 209–228.

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice Hall.

Seth, A. K. (2007). Measuring autonomy by multivariate autoregressive modelling. In F. Almeida e Costa, et al. (Eds.), *Proceedings of the 9th European Conference on Artificial Life (ECAL 2007)* (pp. 475–484). Berlin: Springer Verlag.

Smithers, T. (1995). Are autonomous agents information processing systems? In L. Steels & R. A. Brooks (Eds.), *The artificial life route to artificial intelligence: Building situated embodied agents*. New Haven: Erlbaum.

Thompson, E. (2004). Life and mind: From autopoiesis to neurophenomenology. A tribute to Francisco Varela. *Phenomenology and the Cognitive Sciences*, *3*, 381–398.

Thompson, E. (2007). *Mind in life. Biology, phenomenology, and the sciences of mind.* Cambridge, MA: Harvard University Press.

Van Duijn, M., Keijzer, F. A., & Franken, D. (2006). Principles of minimal cognition: Casting cognition as sensorimotor coordination. *Adaptive Behavior*, *14*, 157–170.

Varela, F. (1979). *Principles of biological autonomy*. New York: Elsevier.

Varela, F., Maturana, H., & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems*, *5*, 187–196.

Weber, A., & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, *1*, 97–125.

Wheeler, M. (2005). *Reconstructing the cognitive world. The next step*. Cambridge, MA: MIT Press.

Wooldridge, W., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review*, *10*, 115–152.

About the Authors

Xabier Barandiaran is a graduate student in Philosophy from the University of Deusto (Bilbao, Spain), he obtained an M.Sc. in evolutionary and adaptive systems at the University of Sussex (Brighton, UK) and received a Ph.D. at the University of the Basque Country (Spain). His main research areas are the philosophy of artificial life, enactive cognitive science and robotics, origins and minimal forms of agency, and naturalized epistemology. He currently holds a post-doctoral grant from the Spanish Ministry of Education to work at Computational Neuroscience and Robotics (University of Sussex).

Ezequiel Di Paolo is a reader in evolutionary and adaptive systems at the University of Sussex and co-director of the Evolutionary and Adaptive Systems (EASy) M.Sc. program. He is a member of the Centre for Computational Neuroscience and Robotics (CCNR) and the Centre for Research in Cognitive Science (COGS). He is the author of over 100 peer-reviewed publications and his interests include: adaptive behavior in natural and artificial systems, biological modeling, evolutionary robotics, embodied cognition, philosophy of mind, and philosophy of biology. *Address*: Dr. Ezequiel Di Paolo, Department of Informatics, University of Sussex, Falmer, Brighton BN1 9QH, UK.
E-mail: ezequiel@sussex.ac.uk

Marieke Rohde received a B.Sc. in cognitive science (Osnabrück, 2003), an M.Sc. in evolutionary and adaptive systems (Sussex, 2004), and a D.Phil. in computer science and artificial intelligence (Sussex, 2008). She currently works as a postdoctoral researcher in the Multisensory Perception and Action group at the Max Planck Institute for Biological Cybernetics in Tübingen. Her research interests include multimodal and sensorimotor integration in humans, time perception and the perception of unity, embodied and dynamical approaches in cognitive science, evolutionary robotics simulation modeling and biological constructivism. *Address*: Multisensory Perception and Action Group, MPI for Biological Cybernetics, Spemannstrasse 41, 72076 Tübingen, Germany.
E-mail: marohde@gmail.com