Available online at www.sciencedirect.com
ScienceDirect
BioSystems 91 (2008) 409–423
BioSystems
www.elsevier.com/locate/biosystems

How (not) to model autonomous behaviour

Ezequiel A. Di Paolo [a,*], Hiroyuki Iizuka [a,b]

[a] *Centre for Computational Neuroscience and Robotics (CCNR), Centre for Research in Cognitive Science (COGS), University of Sussex, Brighton BN1 9QH, UK*
[b] *Department of Media Architecture, Future University-Hakodate, 116-2 Kamedanakano-cho, Hakodate, Hokkaido 041-8655, Japan*

Received 20 February 2007; received in revised form 23 April 2007; accepted 23 May 2007

Abstract

Autonomous systems are the result of self-sustaining processes of constitution of an identity under precarious circumstances. They may transit through different modes of dynamical engagement with their environment, from committed ongoing coping to open susceptibility to external demands. This paper discusses these two statements and presents examples of models of autonomous behaviour using methods in evolutionary robotics. A model of an agent capable of issuing self-instructions demonstrates the fragility of modelling autonomy as a function rather than as a property of a system's organization. An alternative model of behavioural preference based on homeostatic adaptation avoids this problem by establishing a mutual constraining between lower-level processes (neural dynamics and sensorimotor interaction) and higher-level metadynamics (experience-dependent, homeostatic triggering of local plasticity and re-organization). The results of these models are lessons about how strong autonomy should be approached: neither as a function, nor as a matter of external vs. internal determination.

Keywords: Biological autonomy; Modelling autonomous behaviour; Evolutionary robotics; Self-setting of goals; Behavioural preference

1. Introduction

In this paper we would like to establish two important points about autonomy that stem from a careful analysis of the continuity between life and cognition, and a third point by implication. The two main messages we would like to establish about autonomy are: (1) autonomous systems always originate in self-sustaining processes of constitution of an identity under precarious circumstances and (2) such processes can be dynamically manifested in different modes of engaging with the world ranging from committed coping to open susceptibility. The implication of these two points will be that (3) current work in "autonomous" robotics based on ideas of automated synthesis of design (e.g., evolutionary robotics) and dynamical systems approaches to cognition, is still far from achieving or even modelling autonomy in the strong sense advocated here, but that this work may be at the same time the surest route to this goal. We will concentrate for the most part of the paper on discussing examples of recent work in evolutionary robotics. One case illustrates the insufficiency of thinking about autonomy in terms of functions and another example shows that at least some interesting aspects of the organization of autonomous behaviour can be modelled fruitfully once we take points 1 and 2 more seriously. Both cases, however, constitute "good" examples of the role of modelling in clarifying complex concepts such as autonomy.

* Corresponding author.
E-mail addresses: ezequiel@sussex.ac.uk (E.A. Di Paolo), ezca@sacral.c.u-tokyo.ac.jp (H. Iizuka).

0303-2647/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2007.05.016

2. Why should Autonomous Systems Generate their Own Identity?

We will work under the assumption that autonomous systems, i.e., systems capable in some non-trivial sense of setting their own laws, exist, and that living systems provide the clearest, less controversial examples of such autonomy (even if it may still be possible to discuss autonomous systems that are non-living; or let's say, remain agnostic about the possibility). That autonomy is not an illusion is far from evident for Western thought. This is in fact because it is often suspected to be a purely ascriptional property – one that will simply vanish upon closer inspection. Autonomy remains such a slippery concept when examined under the magnifying glass of reductionist physicalism. If we are to avoid mysteries, an autonomous system must follow only the laws of physics, hence it cannot set its own laws, therefore they do not really exist, they are just convenient ways of talking. For Kant, in his *Critique of Judgment*, the intrinsic teleology of organisms was similarly unreachable by pure reason and yet it was so evident that he proposed it should remain as a regulative concept, i.e., we may talk about organisms as if they had purposes of their own but as a convenient shorthand of (not quite well-known) physical events. With autonomy the situation is analogous (and this is no accident). However, the above argument is rather absolutist in its terminological interpretation (what is a law? what is a system?) and its ignoring of the complex possibilities of self-organization of multi-scale physical processes of formation of constraints and structures. More gravely, the argument just too quickly takes sides in the conflict between two kinds of very real experiences: the experience of the physical world as regular and describable in terms of laws and the experience of our perceived teleology and autonomous behaviour in others and, most importantly, in ourselves. On what basis are two reliable and repeatable experiences to be discriminated as real or unreal? History tells us that this is a naive formulation and that conflict breeds novel understanding by dialectical synthesis rather than by decreeing a winner position. This is Hans Jonas's rebuttal of the Kantian lukewarm recognition of the importance, but not quite properly ontological status, of intrinsic teleology. We can know life because we ourselves are alive (Jonas, 1966; Weber and Varela, 2002; Di Paolo, 2005; Di Paolo et al., in press).

Let's just boldly state that living organisms are autonomous – they follow laws set up by their own activity. Fundamentally, they can only be autonomous by virtue of their self-generated identity as distinct entities. A system whose identity is fully specified by a designer and cannot, by means of its own actions, regenerate its own constitution, can only follow the laws contained in its design, no matter how plastic, adaptive, or life-like its performance. In order for a system to generate its own laws it must be able to build itself *at some level of identity*. If a system 'has no say' in defining its own organization, then it is condemned to follow an externally given design like a laid down railtrack. It may be endowed with ways of changing its behaviour depending on history, but at some level it will encounter an externally imposed functional (as opposed to physical) limitation to the extent to which it can change. This can only be avoided if the system's limitations result partly from its own dynamics.

Here we find already a point to be taken seriously by those who pursue the goal of building an autonomous system artificially. It would be wrong to think that the quest for artificial autonomy is futile by definition (to design what cannot be externally constructed). In fact, a subtle change of attitude should take place to start recasting the job of a designer of artificial systems. Once this attitude has changed, there is no contradiction in the idea of strong artificial autonomy. A design process is now transformed into the design of the right conditions (appropriate material substrate and organization) for an autonomous identity to constitute itself. Evolutionary robotics, as we shall see, has made important steps in this novel methodological direction.

The autonomy of a self-constituted system is by no means unconstrained (being able to influence one's own limitations does not imply being able to fully remove them; on the contrary it means being able to set up new ways of constraining one's own actions). Hans Jonas (1966) speaks of life as sustaining a relation of *needful freedom* with respect to its environment. Matter and energy are needed to fuel metabolism. In turn, metabolism sustains its form (its identity) by dynamically disassociating itself from specific material configurations.

Let's provide a definition of an autonomous system.

An *autonomous system* is defined as *a system composed of several processes that actively generate and sustain an identity under precarious conditions*. By *identity* we mean to the joint properties of self-distinction and operational closure. The two properties go hand in hand. *Operational closure*, in a non-trivial sense, indicates the property that among the enabling conditions for any constituent process in the system one will always find one or more other processes in the system (i.e., there are no component processes that are not conditioned by other processes in the network, which does not mean, of course, that other conditions external to the system are not necessary as well for such

processes to exist). *Self-distinction* therefore means the property of a process/component of belonging to such network of enabling conditions (i.e., it is the relation of closure that defines whether a process/component belongs or not to the system), and more strongly, of actively affirming the identity of the system by its own operation. By *precarious* we mean the fact that in the absence of the organization of the system as a network of processes, under otherwise equal physical conditions, isolated component processes would tend to run down or extinguish.

The above definition makes the concept of autonomy *operational*. It should be clear that by expressions like 'self-constitution' and 'generating its own laws' no mysterious vitalism is intended. By saying that a system is self-constituted, we mean that its dynamics generate and sustain an identity. An identity is generated whenever a precarious network of dynamical processes becomes operationally closed. This means that at some level of description, the conditions that sustain any given process in such a network are provided by the operation of the other processes in the network, and that the result of their global activity is an identifiable unity in the same domain or level of description. Autonomy as operational closure is intended to describe self-generated identities at many possible levels (Varela, 1979, 1991, 1997).

For instance, autocatalytic cycles are an example of an operationally closed system in the domain of chemical reactions: by definition, the cycle is capable of sustaining and regenerating itself (given enough supplies) and, at a formal level of description, it defines its own identity: a chemical reaction either belongs or does not belong to an autocatalytic cycle. This identity defines the interactive properties of the system but the history of interaction may also alter the process of continuous identity generation; hence the sense of self-law.

For a living system, the self-identifying processes are themselves processes of material construction and transformation resulting in a self-distinct physical form. The constraint of *physical* construction seems to provide some non-trivial implications to the condition of operational closure. Having a distinct, self-built unity allows to ground consistently notions of behaviour and agency in ways that autocatalytic cycles do not permit. The implication of this is that our definition of autonomy may have to be refined in the future to better grasp the implications of physical self-construction. The notion of precariousness does part of this job. But this issue is not further pursued here.

The definition provided above fits well the case of the constitutive autonomy of living system: their metabolic organization. However, the definition is carefully worded so as to avoid the conclusion that this is the only possible instantiation of an autonomous system. Indeed, we find that several layers of behaviour up to the case of social interactions are able to meet the operational requirements of autonomy (or at least there's no question of principle why they should not). Robotics (the tool for modelling autonomy discussed in this paper) can therefore aim at "catching" the constitutive dynamics of identity generation at some of these higher levels in order to capture forms of non-metabolic generation of values and self-determination (forms that are enabled but underdetermined by metabolism; for detailed discussions on this topic see Jonas, 1966; Di Paolo, 2003, 2005). In other words, we suggest that there are ways of modelling and maybe even instantiating artificial autonomy that do not require building a fully autopoietic artificial system.

In this respect, it is important to indicate that cognitive systems are also autonomous in an interactive sense in terms of their engagement with their environment as *agents* and not simply as systems coupled to other systems (Moreno and Etxeberria, 2005; Di Paolo, 2005). As such, they not only respond to external perturbations in the traditional sense of producing the appropriate action for a given situation, but do in fact *actively regulate* the conditions of their exchange with the environment, and in doing so, they enact a world or cognitive domain.

Viewing cognitive systems as autonomous is to reject the traditional poles of seeing cognition as responding to an environmental stimulus on the one hand, and as satisfying internal demands on the other – both of which subordinate the agent to a role of obedience. It is also to recognize the 'ongoingness' of sensorimotor couplings that lead to patterns of perception and action twinned to the point that the distinction is often dissolved. Autonomous agency goes even further than the recognition of ongoing sensorimotor couplings as dynamical and emphasizes the role of the agent in constructing, organizing, maintaining, and regulating those closed sensorimotor loops. In doing so, the cognizer plays a role in determining what are the laws that it will follow, what is the 'game' that is being played.

The focus on biological autonomy and agency is a radical departure from decades of theories that subordinate cognition to the demands and instructions of either the environment or internal sub-agential modules meant to represent theoretical constructs such as instincts or drives. And it is only made more radical by the connection between the constitutive and interactional aspects of autonomy that is the basis of the idea of sense-making (Varela, 1997; Thompson, 2007; Di Paolo, 2005), the bringing forth of a world of significance.

3. A Fable about the Dynamics of Everyday Life

When trying to understand autonomous *behaviour* it may be instructive to take a look at the ongoing cycles of activity in normal everyday life and how they are often very different from the performances that are studied in psychology, neuroscience, cognitive science and AI/robotics. The locus of study in the majority of work in these disciplines is in general the single performance of an act – the recognition of a pattern, the enactment of a choice, the attainment of a goal, etc. – and the factors and mechanisms involved. It is only rarely the ongoing flow of behaviour that is of concern, i.e., the different modes of engaging with the environment and the autonomous constitution of future engagement such as the emergence of novel goals. In new AI and robotics we find a strong, almost exclusive, emphasis on situated action and *ongoing coping*. This is typically a mode of performance rich in sensorimotor couplings and focused engagement with the task at hand (navigating towards a goal, hammering down a nail, etc.). This mode is highly robust. There are very few distractors that will break down the flow of coping activity. But coping does not always run smoothly. There may be *breakdowns* of different kinds that demand some effort of re-adaptation in order to return to the goal-driven activity. Most of the current work in robotics is about ongoing coping and a fraction of it (dealing with adaptation and learning) is about facing and resolving breakdowns in coping.

Do these two concepts cover all the possibilities that we may encounter in the flow of an animal's (or a human's) everyday activity? This would imply that we lead very busy lives going constantly from one well-defined action to the next, that we are always only coping or dealing with some breakdown and that there is a clear purpose at each moment. If we pay more attention to the temporal organization of goal-seeking coping, we will find that there is one more possible mode of activity. Coping behaviour starts with an intentional demand to fulfill an objective. It is by definition motivated. This motivation or goal is not necessarily fixed or independent of the activity that ensues which is driven by an initial intention or solicitation from the current situation (i.e., a demand or need either external or internal to the agent). However, the fate of all coping, goal-oriented activity is, by its intentional nature, success, abandonment, or frustration (irrecoverable breakdown or simply unattainability). In all cases, coping ceases. We may describe this as the *self-extinction* of all well formed behaviour. If self-extinction does not occur, then we are dealing with compulsive, possibly pathological action (obsessive repetition, moths attracted to the candle flame, etc.).

What happens after self-extinguished coping? It is simply contrary to everyday experience to assume that new goals will immediately follow from the attainment or frustration of previous ones (we are of course not ignoring the possibility of hierarchical organization of tasks into sub-tasks in which case the next set of activities is generally well-defined, but this is not the only possibility). In fact, our experience tells us that there are moments of certain *openness* to the possibilities afforded by our situation (such openness can clearly be very different depending on the affective outcome of the previous coping task). While distractors were robustly ignored during coping, now in an open state with undefined goals, an agent may be drawn by environmental or internal events into forming a novel intention and retroactively investing such a "distracting" event with meaning (for instance, I decide to put down the page I am reading, I take a deep breath and look around my desk aimlessly, the sound of a car horn in the distance makes me look out of the window and on seeing the garden I notice that some maintenance is now long overdue, I decide to go and do some work there, it was one of the things I was previously intending to do, but not just now, the car horn "reminded" me of it). So there are durable states of dynamical openness and susceptibility to micro-events that are qualitatively very different from intentional, goal-driven coping. Openness does not self-extinguish by the logical structure of an intentional act, but it is bound to be extinguished by its very nature of high susceptibility.

The different modes are represented in Fig. 1. Dynamically speaking, we could venture the hypothesis that coping relates to low-dimensional, highly robust, coordinated body/environment dynamics whereas openness relates to high-dimensional, typically unstable, uncorre-

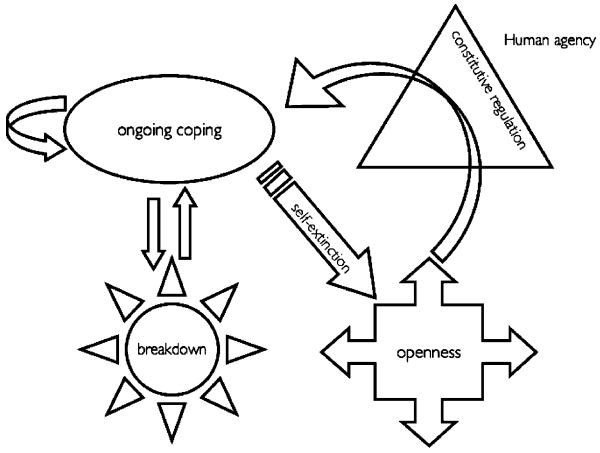

Fig. 1. Dynamical modes describing the flow of everyday activity.

lated dynamical modes. And that switching between one and the other is not symmetrical. And yet the emerging picture is one of transiting between very different dynamical modes, between stability and instability. Maybe, in order to be able to synthesize artificial autonomous systems, we must first understand better what sort of dynamical system can generate such different regimes.

There is a further step in the story of transiting between different modes of activity that is well supported by phenomenological analysis for the case of human agency. This is the active and regulated *constitution* of goals. The path from a state of openness to new coping simply just happens for most animals. The new goal does not wait long before it appears. But humans can bring about a recursive constitutive skill to this passage by the act of simply asking: *And now what? What was it that I had planned to do?* Therefore, a further level of autonomy that is supported by socio-linguistic skills and that down-regulate behaviour by actively constituting new goals marks human agency. However, this higher form of autonomy is beyond the scope of the present paper.

4. Limitations of Pure Evolutionary Robotics

Let us now turn to the problem of modelling autonomous behaviour. Let us immediately clarify the obvious but easily forgotten fact that models and instantiations are very different things. An instantiation is supposed to be a proper member of the class of phenomena under study. By contrast, a model need not be; it can be crude – almost ignoring the majority of interested aspects of the phenomenon of interest – and yet be extremely useful. In general, simple models tend to be scientifically very powerful. This is because the purpose of a model is *not to replicate* a phenomenon, but to help *explain it*. There are lots of ways in which this can happen that do not involve producing an instantiation: models can show us the mistake in our assumptions, they can be explanatory rich in the way they actually fail to capture the phenomenon of interest, they can act as proofs of concept, they can generate novel hypotheses, and generally they can help re-organize complex ideas by exercising and questioning our intuitions. The models on self-generation of goals discussed later in this paper help us think about the concept of strong autonomy discussed in the previous section without either of them coming close to being an instantiation of this concept.

Modelling (and instantiating) autonomous behaviour is the goal of robotics. However, in robotics the term autonomous is often used very loosely. It can mean anything from mobile, un-tethered, adaptive, to self-recharging or self-powered. In the sense of not being constantly controlled from the outside and being able to cope with a noisy, real-world environment, mobile robots imitating simple lifeforms have been seriously investigated for the last two decades (Brooks, 1991) and have noble ancestors such as W. Grey Walter's tortoises (Walter, 1950). Such robots move about using simple but powerful principles of engaged interaction and achieve robust performance in the absence of explicit controlling at the level of attaining a certain goal. Robust performance emerges from the interaction of simple mechanisms with body and environmental dynamics. These robots exploit loose couplings with the environment to achieve sustained behaviour. But what about autonomous agents in a sense that is closer to the autonomy of living systems, agents capable of setting their own laws? As argued above, as long as a system is externally designed (even in terms of eventual changes that it may undergo in its organization) and not allowed to constitute itself, it cannot really be autonomous in the strong sense. Its goals are not set by itself but by the designer; they are extrinsic to it. However, interesting behaviour that approaches different aspects of autonomy is often observed once the designer starts constraining the process of design at increasingly removed levels.

Evolutionary robotics (ER) is still proving a useful and open-ended method for exploring this increasingly less constraining role of the designer that may be required to achieve strong artificial autonomy. ER hands in the task of filling in design specifications pertaining to mechanisms, morphology, structural and functional organization to an automatic process of artificial evolution (Harvey et al., 1997; Nolfi and Floreano, 2000). Thus, instead of designing a robot that must explore the environment but should go to the green light when the battery is down, one can attempt to design a robot that more generally must keep the battery up during its explorations, or more implicitly, a robot that explores indefinitely. In principle, there may be different ways of achieving this broader goal, and artificial evolution can find many of these ways and select for robots that opportunistically choose the most convenient route towards the general goal of maintaining the battery charged (it may imply taking advantage of a different source of energy that the one we intended as designers). A proper, strongly autonomous agent (in a sufficiently complex environment affording many alternative routes towards a goal) would certainly maximize the selection criteria. So, one could hope, all that needs to be done is evolve such robots for sufficiently long times and such autonomous agents will emerge eventually.

Unfortunately, this is too optimistic a view and relies on a misunderstanding about artificial evolution. Unlike

the open-endedness of natural evolution (operating on systems that are already autonomous), artificial evolution tends to be conservative rather than innovative. Or, rather, its innovation resides in that it often finds simpler, cleverer solutions than the ones we expect as designers. This is what makes it a powerful scientific tool to debunk myths and clarify pre-conceptions (Harvey et al., 2005). Artificial evolution is capable of producing such results because it works outside the box of design constraints that limit the way we think about the system and the task it must achieve. But as a process operating on statistical information about a set of tested solutions to a problem, it will always run the risk of getting stuck on solutions that are statistically mediocre and thus finding it hard to explore elements of design that are initially neutral but that allow novel possibilities if they are reliably present.

The best way to illustrate this is with an example. Tuci et al. (2003) have investigated landmark learning behaviour in a mobile, Khepera-like, robot controlled by a continuous-time, recurrent neural network (CTRNN) without synaptic plasticity. That learning is afforded without changes to a network's connectivity is already one major demonstration of the power of ER to break pre-conceptions. This work intended to reproduce previous work by Yamauchi and Beer (1994) on landmark learning in one dimension, but with an agent moving freely in a two dimensional arena. The task is simple: find the location of a goal that cannot be seen from the distance using information provided by a fixed light. In half the cases, the light is next to the goal, in the other half the light is far from the goal. Approaching the light and remembering the relation to the goal would enable an agent to learn in which of the conditions it finds itself, and then on later trials move either towards the light or away from it, but always towards the goal. Yamauchi and Beer had to use a modular de-composition of functions into sub-networks to solve this task. It was the suspicion of Tuci et al. that this was unnecessary. However, if the fitness function were simply to count how many times the agent found the goal after the first trial (minimising designer involvement), no learning would evolve. Effectively, evolution settles on a fixed strategy (e.g., always move left) and finds the goal 50% of the times – a result that will not be unfamiliar to practitioners of ER. Why is the robot not using the light? Because the light is *on average* uncorrelated with the goal position, and fitness is measured as the average of many trials with the same number of presentations in each of the two possible situations (landmark far, landmark near). Hence, the safest bet given this lack of correlation is that the light is a distractor and hence it should be ignored.

Tuci et al., realizing this, solved this problem by introducing an artificial bias in the selection process (effectively becoming more involved in it as designers). During the initial phase of the evolutionary process, they gave extra rewards to robots that approached the light (on top of whether they also approached the goal or not). This forced the evolutionary process to select neurocontrollers that responded to the light as a relevant stimulus. When this extra fitness criterion was removed in later generations (and only goal seeking remained) the population consisted of neurocontrollers that firstly sought the light, and from that situation they had to work out what to do next. The light, because of the initial bias, has ceased to be irrelevant, and in such circumstances (standing on the same spot as the light after having approached it) now the evolutionary process can uncover the proper correlation between light and goal. As a result of this, robots capable of learning the landmark correlation to a goal evolved.

So we may conclude that more sophisticated levels of behaviour in general (as more sophisticated models of autonomy) may demand more and not less design intervention in the evolutionary process. *This is the apparent paradox of artificial autonomy*. The system should in some sense build itself, the designer should intervene less, but it should at the same time be more intelligently involved in setting the right processes in motion. Contrary to the uninformed perceptions at the time when ER was born, one cannot treat artificial evolution as a magic box capable of solving any problem one poses to it (and all one must do is just wait). Fortunately, failures to evolve a desired behaviour, if followed by some analysis of the behaviours that do evolve, often leads to a revelation of what are the problems one must overcome as a designer of an evolutionary regime.

In addition, by its very nature, ER proceeds by testing candidate solutions under a set of varying circumstances in order to select robot controllers capable of latching onto the *significant interactions* with the environment that will lead to achieving the desired goal efficiently and robustly. Finding a target cannot depend on the initial position of the agent, or the initial internal state, and so these parameters must be randomized from trial to trial to ascertain a level of stability of the solutions that evolve. But this very basic element of the ER methodology may play against the design of autonomous agents, at least if we consider the different dynamical regimes of activity described in the previous section. If evolution is to produce stable and robust dynamical controllers, it will avoid being strongly influenced by irrelevant environmental factors, but at the same time it will avoid internal sources of instability. Hence, it will produce robust cop-

ing, but not necessarily dynamical states of openness after coping activity is self-extinguished. That's why goal-seeking evolved robots tend to keep around their targets like moths attracted to a flame. Their behaviour is almost pathological. The lack of self-extinction of behaviour should perhaps be taken as a sign of bad design (cf., work by Ian Macinnes on functional circles and practical ways of dealing with this problem, e.g., Macinnes and Di Paolo, 2006). So evolving autonomous robots will have to overcome this problem by either selecting the right building blocks, or including sensorimotor interactions and internal elements that inevitably will sometimes lead to transitions between low and high dimensionality in the dynamical flow as suggested in the previous section.

5. A "Self-instructing" Agent. How not to Model Autonomy

Let us consider an example of an agent capable of generating its own instructions and following them. In some loose sense of autonomy (but not necessarily in the operational sense that we have offered above), this agent would be setting up its own goals. We present the following agent as a computer-enhanced thought experiment but also as a demonstration of why certain tempting methodologies for designing autonomous agents are conceptually flawed. In the next section, we will show an agent that is not yet fully autonomous but which demonstrates what we consider a better methodology. Both these models demonstrate how we can learn about autonomy without yet producing proper instantiations.

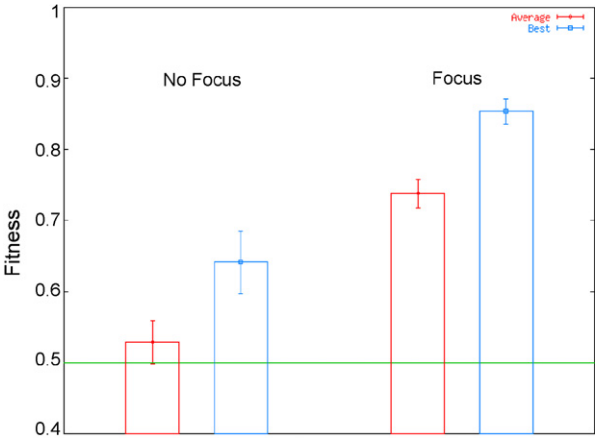

Fig. 2. Fitness achieved for agents evolved to catch circles or diamonds under external instruction with and without focus control of sensor rays. Average of 10 independent runs.

In his well-known discrimination experiments, Beer (2003) has shown how minimally cognitive behaviour can be (1) easily modelled and analysed using a combined evolutionary robotics/dynamical systems approach, and (2) how such models, albeit minimal, demonstrate interesting general principles and provide extendable vocabularies to discuss cognition in dynamical terms. The basic discrimination experiment consists of a visually-guided agent moving in 1 dimension (left-right) whose task is to catch a falling object if it is a circle and avoid it if it is a diamond; the agent receives input from an array of linear visual sensors (rays that activate when intersected by the falling object) and this input is fed into a recurrent, symmetrical CTRNN controller.[1] The output of the network determines the velocity of the agent (Beer, 2003). Although dynamical analysis has shown that agents use the absolute radius of the falling shape to perform their discrimination, extensions of the setup to shapes of variable size results in agents capable of discrimination based on shape, (Di Paolo and Harvey, 2003).

Let us consider a variant of this model. An agent that performs a circle/diamond shape discrimination, but that depending on an external binary signal its choice of which object to catch can be altered. So if the external signal (ES) is set to 0, the agent is a circle-catcher and if the signal is 1, the agent is a diamond-catcher.

The setup is otherwise similar to Beer's experiments, with the difference that sensors are binary (to increase sensory ambiguity and encourage more active solutions). And additionally, a focus control is added to the array of sensor rays. This is an effector neuron that simply opens and closes the angle of the sensors rays in a linear way. Interestingly, this extra level of sensory control is important to evolve agents capable of changing their behaviour depending on the external instruction. Fig. 2 shows the average fitness of 10 independent runs with and without focus control. The best focus controlling

[1] The state equation for a CTRNN neuron is:

$$\tau_i \left(\frac{dy_i}{dt} \right) = v_i + \sum_j w_{ji} z_j + I_i,$$

where i indexes all neurons, j indexes all links inputting to neuron i (which may be an empty set), τ_i is a time constant, y_i is the neuron state (analogous to a membrane potential), I_i is an input current, w_{ji} is the link weight from neuron j into neuron i, and z_j is the activation of the pre-synaptic neuron attached to link j. For a neuron, the firing rate is given by the logistic function:

$$z_j = \sigma(y_j + b_j) = \frac{1}{\{1 + \exp[-(y_j + b_j)]\}},$$

where b_j is a bias parameter.

agents can perform either circle-catching or diamond-catching on demand for a relatively large range of sizes, using ambiguous noisy sensors with success rates of over 85%.

These agents have now a well-defined signal that alters the goal they pursuit. Could not such a signal be somehow provided internally? Ideally, could such a signal be generated in a way that is jointly dependent on internal and environmental factors? Exclusive dependence on either class of factors would not generate an agent that we would be happy to call autonomous as we could suspect that the agent is following the instructions that either are external to it or is blindly taking no account of its situation. Autonomy, even in an intuitive sense, is ruled out by either of these two conditions. Why? Because both conditions negate the idea of self-determination. The case of constant reactive response to the environment is clear. No system that is simply driven externally can ever be autonomous. But, and this is less intuitive, the same may be said about a system that is "driven internally". If a subset of a system exerts control on the whole, then the situation remains that of a system that is *controlled*, not *self-determined*. If a system is controlled only by internal dynamics making it blind to the current environmental situation (what sometimes in mathematical terms is indeed called an "autonomous" system due to the lack of parametrical and time-dependent driving), the system has nothing to determine itself *against*. It simply endures in its dynamics because it is closed to environmental challenges.[2]

So taking these intuitive constraints into account let us conceive of a sub-system capable of generating a stable on-off signal depending on internal state and environmental circumstances. There are many options. One would be a central pattern generator (CPG) that oscillates with a certain frequency in the absence of input currents and settles into either a high or a low value stable attractor in the presence of input. Such a circuit can easily be hand-designed using a fully connected 2-node CTRNN (Beer, 1995) and is shown in Fig. 3. The CPG receives input from the visual sensors. Depending on the phase value of the oscillation orbit, the presence of input will drive the CPG to one of two possible stable fixed points (new intersections of nullclines[3]). For one of the nodes the two fixed points correspond to high and to low firing rates respectively. This node is then connected to ES in the pre-evolved discriminator network. The agent will now produce behaviours such as those shown in Fig. 4. Upon repeated presentation of a circle the agent will sometimes approach it and other times avoid it. Similarly for diamonds. In a very simplistic way, the agent is setting up its own "goals" by instructing itself to go for one object type or the other. It does so in a way that depends on internal conditions (phase in the CPG cycle) and external factors (e.g., position and timing of the falling object). An external observer could describe the agent's behaviour as "capricious".

Why is this a problematic way of approaching artificial autonomy? Even though this model tries to capture an intuitive notion of autonomy as the setting of a system's own goals by the system itself, the integrity of what we take to be the system ultimately relies on the designer's viewpoint. It is (like most artificial agents to this day) *a system by convention*. It is only too clear that the add-on of a CPG to an already evolved neural network does not result in a system that is integrated in other than a nominal sense. In fact, because they system's identity is given externally, we could just as well visualize the CPG circuit as located in another room and communicating with the system through remote control. And this is in effect what a method aiming at integrating a *controller* into the system itself (see Smithers, 1997) will always achieve: a homuncular solution whereby the human controller is replaced by a module telling the rest of the system what to do. The advantage of this example is the clarity with which this problem presents itself, but more sophisticated versions of the same idea (when clear modular separation is not so easy to perform) will be conceptually not different. Hence, proper autonomy must address internal goal generation as a result of the system's own organization, rather than as a function that the system produces (independently of whether such

[2] If such blind action were to be the paradigmatic case of autonomy, we should think of mountains as being alive since they endure much longer than living systems. But life is not about enduring and autonomy is not about blindly ignoring the environment. Self-determination becomes an empty concept if the system is detached from sources of uncertainty and solicitations that would tend to induce in it alternative outcomes from the one that the system itself is struggling to achieve. In this view, autonomy is always a dialectical concept.

[3] Nullclines in a 2-node CTRNN circuit are calculated by setting the derivatives of the states y_1 and y_2 equal to zero in the CTRNN state equation. The y_1 nullcline is given by:

$$J_2 = \ln\left[\frac{J_1}{(w_{21} - J_1)}\right] - \frac{w_{22}J_1}{w_{21} - I_2 - b_2}.$$

The y_2 nullcline is given by:

$$J_1 = \ln\left[\frac{J_2}{(w_{12} - J_2)}\right] - \frac{w_{11}J_2}{w_{12} - I_1 - b_1}.$$

where $J_1 \equiv w_{21}\sigma(y_2 + b_2)$ and $J_2 \equiv w_{12}\sigma(y_1 + b_1)$. See (Beer, 1995).

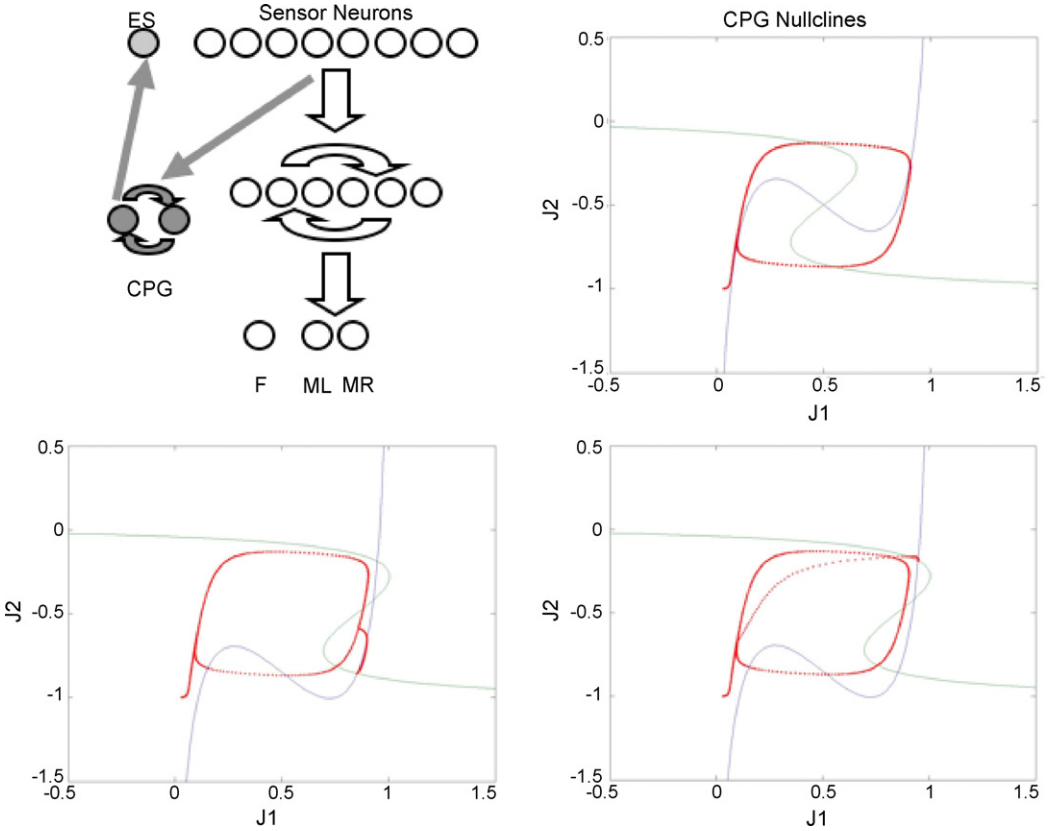

Fig. 3. Top left: CTRNN neurocontroller for self-instructing agent. ES: external signal, F: focus effector, ML and MR: motor neurons driving left and right, respectively. Top right: nullclines corresponding to fully connected 2-node CPG in the absence of input. Bottom left and right: nullclines in the presence of input, trajectory ends in a low firing fixed point for neuron 1 (left) or in a high firing fixed point (right) depending on phase. Output of neuron 1 is fed into ES.

function is achieved in a modular or distributed manner). In short, *autonomy is not a function*. A similar point is made by Rohde and Di Paolo (2006) with respect to value generation and value systems (see also Di Paolo et al., in press; Rutkowska, 1997). We are faced with an important consequence of the view on biological autonomy sketched at the beginning which is not necessarily obvious at the moment of starting to develop it into a workable model: autonomy is not something that a system does, it is a property of how the system is organized and re-organizes itself so as to channel its functionalities towards newly generated intentions.

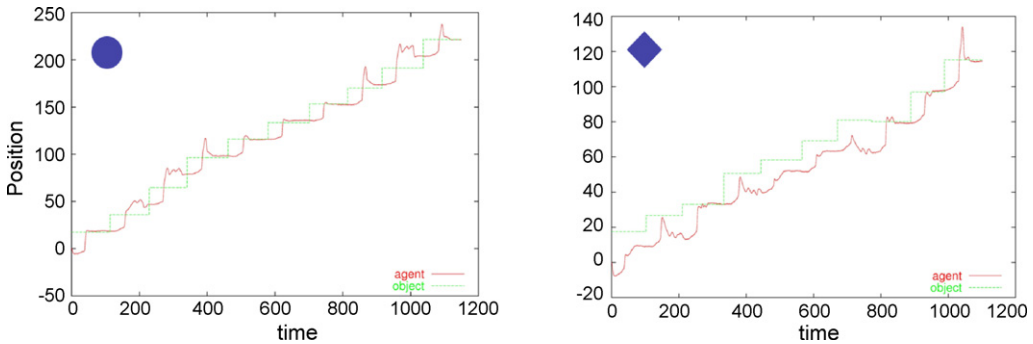

Fig. 4. Repeated presentation of falling circles (left) and diamonds (right) for self-instructing agent. Plots show the horizontal displacement of the agent over time and the position where the objects fall. Agent sometimes approaches the target, other times avoids it.

6. A Better Way: Evolving Behavioural Preferences

For ease of comparison, let us stay with the problem of an agent producing an autonomous behavioural choice, but let us approach the problem in terms of the dynamical organization of the agent. How would such a model look like? What sort of method can produce a system that by its very structure and interactions would by itself select and regulate a sensorimotor flow based on the effects it has on the maintenance of an internal organization?

Consider a dynamical model of the formation and sustaining of a behavioural preference. How does an embodied agent develop a stable preference such as a habit of movement, a certain posture, or a predilection for spicy dishes? Is this development largely driven by a history of environmental contingencies or is it endogenously generated? Goldstein (1934) described preferred behaviour as the realization of a reduced subset of all the possible performances available to an organism (in motility, perception, posture, etc.) that are characterized by a feeling of comfort and correctness as a contrast to non-preferred behaviour which is often difficult and clumsy. In this view, the fact that a preferred behaviour is observed more often would be derivative of these properties and not central to its definition (preferred behaviour is often efficient but not necessarily optimal in any objective sense). Following this idea, a preference can be defined as the enacting of a behavioural choice that is sustained through time without necessarily being fully invariant, i.e., with time it may develop or it may be transformed into a different preference.

Behavioural preferences and their changes lie between the two scales typically addressed by dynamical systems approaches to cognition: the behavioural (e.g., Beer, 2003; Kelso, 1995) and the developmental (Thelen and Smith, 1994) and share properties with both of them. Goldstein has argued that we cannot really find the originating factors of a preferred behaviour purely in central or purely in peripheral processes, but that both the organism's internal dynamics and its whole situation participate in determining preferences (Goldstein, 1934). In this view, it becomes clear that a preference is never going to be captured if it is modelled as an internal variable (typically a module called "Motivation") as in traditional and many modern approaches, but that a dynamical model needs to encapsulate the mutual constraining between higher levels of sensorimotor performance, and lower processes, such as neural dynamics (Varela and Thompson, 2003). This is what the above model does not achieve. A preference is not "located" anywhere in the agent's cognitive architecture, but it is rather a constraining of behaviour (through internal and external conditions) that is in turn shaped by behaviour.

Iizuka and Di Paolo (in press) present an exploratory model of behavioural preference with the objective of exploring Goldstein's assertion of multi-causality. This model illustrates a potentially fruitful method for modelling autonomy as well. The minimal requirement to capture the phenomenon of preference is a situation with two mutually exclusive options of behavioural choice. An agent should be able to perform either of these options but the choice should not be random, but stable, and durable. The choice should not be invariant either but it should eventually switch (in order to study the factors that contribute to switching). There should be a correspondence between internal dynamical modes and different aspects of behaviour (from commitment to a choice to switching between choices). For this the methods of homeostatic adaptation (first introduced in Di Paolo, 2000) are used to design not only the agent's performance but to put additional requirements on the corresponding internal dynamics.

In the original model of homeostatic adaptation an agent is evolved to simultaneously perform a task and to maintain its internal variables (e.g., neuronal firing rates) within some homeostatic bounds, (Di Paolo, 2000). When such variables cross the boundary, local internal plasticity is activated, and keeps active until homeostasis is recovered (Ashby, 1960). Some of the agents that evolve under these conditions show a dynamical link between the two objectives (performance and internal homeostasis) such that disruption to performance (e.g., changes to motors or sensors) result in internal instability, which provokes plastic internal changes until stability is regained. Because of the dynamical link established during evolution, regaining internal stability involves a behavioural adaptation to the original disruption. The result is that the agent is able adapt to severe disruptions that have not been presented to it before.

This idea can be naturally extended to a situation where an agent must choose between alternative behaviours: A and B. Instead of a single homeostatic region for the internal variables (neuronal firing rates), there are two regions. If the neural dynamics are contained within either of these high-dimensional boxes, the network remains unchanged. But if the flow of internal states moves out of the boxes, local plastic synaptic changes are applied in the general direction of reverting the flow to move again inside the box. Now, agents are evolved to perform each behaviour A and B and to behave homeostatically so that the internal state is inside box A while performing behaviour A and inside box B while performing behaviour B. The evolutionary regime

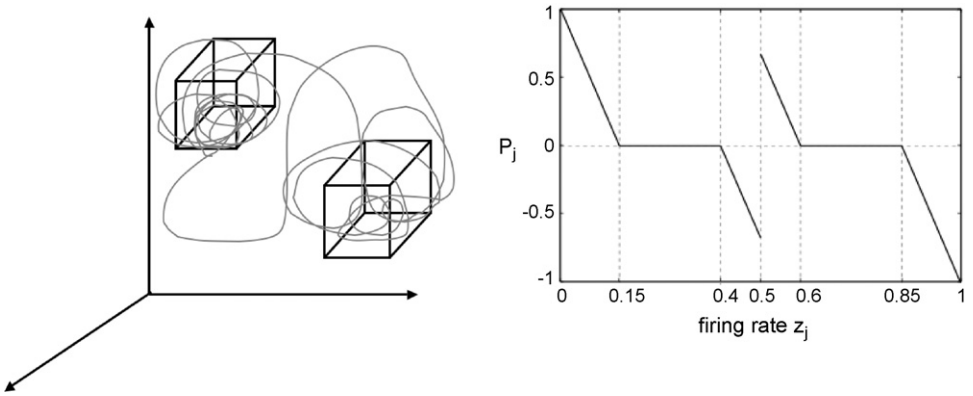

Fig. 5. Left: schematic representation of two high-dimensional homeostatic regions in the space of neural firing rates. Right: how the homeostatic regions are implemented for each node in the network. The plot indicates the plasticity function (p_j) as a function of neural firing rate (z_j). Changes to incoming weights are calculated as a function of pre- and post-synaptic activation multiplied by p_j. Whenever the post-synaptic firing rate is in one of the two flat regions, $p_j = 0$ and local plasticity is inhibited.

is designed so that behaviour is not biased to either of the choices. In the case presented in (Iizuka and Di Paolo, in press) the behaviour is simply approaching a light of type A or B (the wheeled Khepera-like agent has two pairs of sensors left and right towards the front, one for each type of light). Each homeostatic box is implemented for each node in the CTRNN neurocontroller as 2 bands within the range of firing rates (Fig. 5): a low-firing and a high-firing homeostatic region (to reduce bias, the type of each region, A or B, is assigned randomly at the beginning of the evolutionary run).

The idea is that if the system holds two separate (fixed) high-dimensional boxes in the space of neural dynamics which are associated with performing different behaviours, a preference could be formed by the dynamical transitions that select which box the trajectories go into and stay in. This provides a first requirement for talking about preference, that of *durability* (bottom-up construction of stability). Once a behaviour is formed, due to the stability in a box, the system keeps doing the behaviour while ignoring other behavioural possibilities. It is like a spontaneous top-down constraint that regulates the sensorimotor flow. However, some disturbances might eventually cause a breakdown of the stability and then another behaviour can be reconstructed though the homeostatic adaptive mechanisms. Since by design, the system has another region of high stability, it will be possible in the right circumstances to switch into it and then start enacting the other behavioural option. In this way, behaviour can switch due to the corresponding transitions between two boxes. One can expect to see both spontaneous and externally induced transitions from the viewpoint of the top-down and bottom-up construction or destruction of durable but impermanent dynamical modes. Here, we find a second requirement, that of the possibility of *transformation*, or change in preference.

The evolved agents show interesting behaviour when two lights (A and B) are presented simultaneously in a random position. They always "chose" to go to one of the two lights, they never stay in the middle or move away from them. Moreover, when the position of the light is replaced by two new distant positions, the agents seem to maintain a preference for the light they have visited previously, but not indefinitely. Fig. 6 shows a sequence of 100 consecutive presentations of lights (A and B) in sequence with randomized positions. The plot on the left shows the final distance to each type of light. We can see that the agent approaches light B for the first 25 presentations, but then switches to light A and maintains this behaviour for about 35 presentations before switching again. The plot on the right shows the corresponding proportion of time that the neural dynamics is inside boxes A and B. It is clear that the proportion tends to be high while the agent is performing the corresponding behaviour.

Many tests have been performed to assess what makes an agent change its preference, (more details in Iizuka and Di Paolo, in press). For instance, while approaching light A, the lights are swapped in position to see whether the agent changes its behaviour. The result depends on the time of the swapping. If the agent is far enough, it alters its trajectory after the swap and moves towards the new position for light A. If the swap is made later, when the agent is close to light B, the agent switches to finish its approach to light B, as if its presence was now too strong a stimulus to ignore. This and similar tests indicate that a preference is maintained or changed as a combined effect of environmental factors and endogenous dynamics.

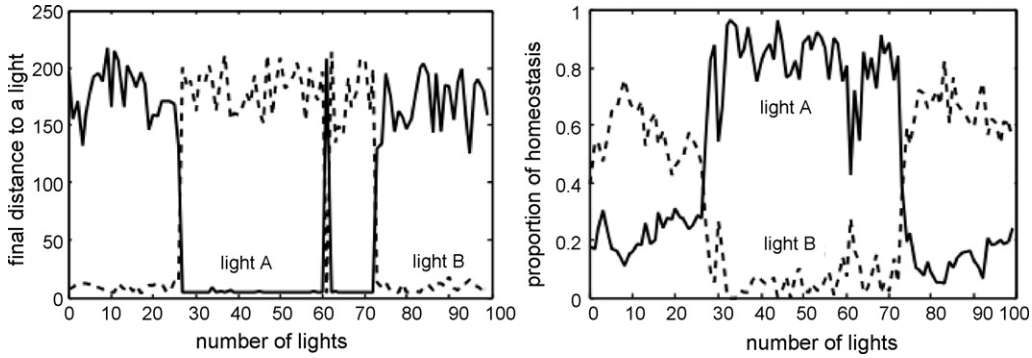

Fig. 6. Left: final distance to each light at the end of trials on serial presentations of 100 pairs of lights. Right: proportion of neurons that have stayed within the homeostatic region for each light in correspondence to trials on the left. Adapted from Iizuka and Di Paolo (in press).

In an attempt to measure the development of a preference, agents are tested at different times during the sequence of presentations shown in Fig. 6 in order to find out if their choice would have been the same at that point in time if the position of the lights had been different. The distinction between a spontaneous or externally driven "decision" is made operational by observing the agent's behaviours in different situations departing from a same initial state. If the agent "decided" to go to one of the lights endogenously, its behaviour must be robust without depending too much on environmental factors. On the contrary, if the selection were externally driven it would be affected by changes to environmental factors such as light positions (as if the agent were not "committed" enough).

Fig. 7 shows the results. Each plot indicates in shades of grey the final destination approached by the agent as a function of the angle relative to the body in which each light is presented. The neural and bodily states are picked from those corresponding to a given time in the sequence shown in Fig. 6. In the case of (a), in which the agent originally has the preference of light B, the "decision" is stable against the various alternative positions of the lights. The agent robustly approaches light

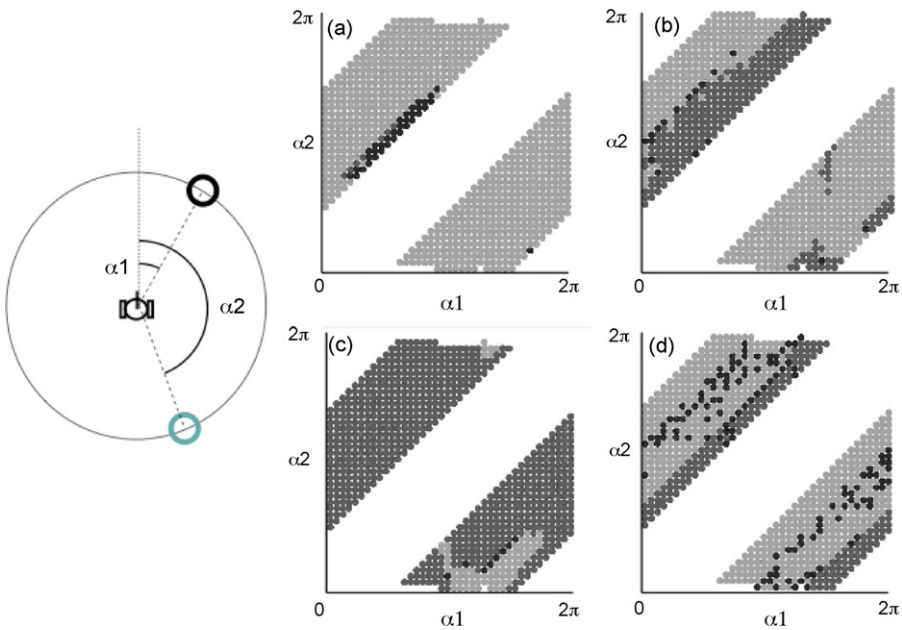

Fig. 7. Light preference of the agent corresponding to the states of (a) 20, (b) 25, (c) 50 or (d) 95 in Fig. 6, against different light positions. Horizontal and vertical axes indicate the initial angles of lights A and B relative to the agent's orientation, respectively. The positions of lights whose difference is less than $\pi/2$ are removed in order to better determine which light the agent is approaching. The dark grey circles show that the agent approaches light A. The light grey circles correspond to light B and black shows the agent does not approach either of lights. Adapted from Iizuka and Di Paolo (in press).

B for practically all the angular positions tested. Therefore, the "decision" to approach B does not strongly depend on environmental factors in this case. This is also true in the case of (c). Except for the small region where it selects light B, the agent approaches light A wherever else the lights are placed. By contrast, in cases (b) and (d) the agent is rather "uncommitted" because the approached target changes depending on the lights' position.

The agent thus alternates between periods of high and low preference for each of the options presented to it. During periods of high preference it could be said that the agent's behaviour is committed in the sense that it is less dependent on distracting factors (as implied in Fig. 7, the agent will under these circumstances ignore the "wrong" light even if placed directly in front of it, and look for the "right" light even if it is out of its visual range). While the preference is changing from one option to the other the agent does not show a strong commitment to which light should be selected as target. There is ample scope for distracting factors to alter the agent's behaviour. Then, the new preference develops towards a more stable (or committed) dynamics. It is this *alternation* between these modes – resembling the alternation between the modes of *coping* and *openness* described in Section 3 – that makes this behaviour closer to something we might call autonomous.

The implication of these results is that the emergence of a new goal happens by an interplay between internal and interactive dynamics and that it consists, not of a function that performs certain decisions and instructs how the agent should behave, but of the regulation of periods of openness to external and internal variations and periods of commitment to a goal. Even during periods of weak environmental dependence, the endogenous dynamics are not solely responsible for the agent's performance. In all cases, behaviour is the outcome of a tightly coupled sensorimotor loop. It is clear that the mode of environmental influence, whether weak or strong, changes over time and that this is a property of the agent's own internal dynamics as well as its history of interaction. During the periods of high susceptibility to external variations, the agent is highly responsive to environmental variability resulting in less commitment towards a given target. By contrast, during periods of weak susceptibility, the consistent selection of a target is a consequence of low responsiveness to environmental variability.

The important point is that the autonomy of the agent's behaviour can be seen as the flow of alternating high and low susceptibility as suggested in Section 3, which is an emergent property of the homeostatic mechanism in this case (but might be the result of other mechanisms in general). There is nothing apart from the flow of neural and sensorimotor dynamics that stands for a mode of commitment to a preference or other. It should be made clear that this picture is quite a contrast with the idea that autonomy may be simply measured as how much of behaviour is determined internally versus how much is externally-driven. Strong autonomy is orthogonal to this issue since simply all of behaviour is conditioned by both internal and external factors at all times. It is the mode of responsiveness to variations in such factors that can be described as committed or open, and it would be a property of strong autonomous systems that they can transit between these modes (maybe in less contingent ways as this agent).

This agent comes closer to some properties of autonomous behaviour described above, especially in addressing the constitution of a new goal as an emerging property of the internal and interactive dynamics in relation to organizational constraints. We cannot really claim that the agent is fully autonomous however. The fixed internal constraints (homeostatic boxes) are rather arbitrary in their definition and their lack of contingent development over time. It seems that having homeostatic regions that are somehow themselves constituted by a history of interactions would be a much better way of modelling autonomy. In addition, there is only a contingent link between internal "requirements" and external interaction. Light is made relevant to the agent by a selective pressure and it is linked to an internal condition to be satisfied (homeostasis) also by evolutionary history. Organisms present much tighter double causal links between internal needs (e.g., metabolism) and sensorimotor interactions (e.g., foraging). This is again something that should be improved for a closer approach to behavioural autonomy, (Di Paolo, 2003).

7. Conclusions

What do we learn from these models? The two examples of modelling aspects of autonomous behaviour presented above allow us to draw some clear conclusions that are implicit but not well articulated in the more conceptual view on biological autonomy proposed in Section 2: Autonomy (and its implications such as identity generation, value-generation, goal-setting, etc.) is an organizational property of a system, not a function, a state or a mechanism. Any attempt at approaching it purely in functional terms will miss something fundamental.

In other words, autonomy is a property pertaining to what a system *is*, rather than what it *does*. This onto-

logical property will have very clear consequences for how the systems behaves, i.e., what it does. But to start a model from those consequences will always run the risk of trivializing autonomy or even explaining it away. In our two cases, we have focused on the self-setting of goals as an example of an autonomous performance. Implementing this in an intuitive manner results in a model that is not satisfactory because it misses the ontological dimension of autonomy. It treats it as a function. A more sophisticated implementation is able to capture the underlying dialectics between dynamics and meta-dynamics (homeostatic constraints and plasticity), between organization and performance, and goal setting is achieved in an emergent manner and not as a function of the system requiring some dedicated computational module. There is simply nothing in the neural controller of the agent that sets out new goals, and yet this is what the agent does as a whole.

The preference model also suggests a second lesson. It is not very fruitful to ask of a putative autonomous system whether its behaviour is caused by internal or external factors. This approach breeds confusion because all of behaviour is always determined by both internal and external conditions. Autonomous behaviour is, like the preferences shown by the agent, always caused by a multiplicity of internal and external factors. It is the response of the system to variability in such factors that gives an idea of the particular mode of commitment to a goal, and it is to be expected of autonomous systems that they would also show transitions between these modes as shown in the preference model. The study of this model has so far looked at the commitment to a preferred behaviour by altering environmental conditions (positions of light), but we might as well study internal variability (e.g., noise or lesions) such that behaviour is still achievable and similarly measure different modes of engagement resulting in analogous periods of coping and switching (this study has not been performed yet).

Are we dealing with an autonomous system in the model of preference formation? Not yet. We may ask (as suggested at the end of the last section) to what extent is an identity self-generated in this system. This definitional aspect of autonomy is not captured by this model. It is clear that much of the system is rather stable and not precarious (e.g., the agent's body, sensor and motor responses, and neural connectivity) and that if an identity could be self-generated anywhere in this case, it would have to be at the level of the combined neural and sensorimotor dynamics. But still there is much arbitrariness in the design of this setup (such as the location and static nature of the homeostatic regions). In fact, whether higher and recursive levels of identity are possible without the grounding in physical self-construction is still an open question.

The complex dialectics between different dynamical levels is at the root of several intentional aspects of autonomous behaviour from the generation and appreciation of values, norms and affect in a situation, to the emergence of a sense of agency and self. We expect that, by producing models that either make confused ways of thinking more manifest or indicate more clearly the relation between complex ideas, this kind of methodology will help us further research into these related questions. At the moment, functional modelling (attempting to capture concepts such as autonomy or agency in terms of functions) is still prevalent (Di Paolo et al., in press; Rohde and Di Paolo, 2006). This is in part due to the technical limitations of traditional approaches to cognitive modelling (which we are beginning to overcome) but also partly due to some conceptual Cartesian baggage that hides itself under apparently innocuous assumptions, especially in terms of deriving mechanisms for externally observed functionality.

However, we must at all times remember the point of producing models such as the ones described here. They are not meant to be implementations of the properties under study. The formation of an identity may well be modelled at the neural level (e.g., the formation of a dynamic pattern that is self-sustaining under precarious conditions and whose maintenance require certain sensorimotor interactions with the environment) even though implementing such a process may not still be enough to implement a proper autonomous agent. A model is supposed to expose gaps in our understanding – not produce fancy performances. For this, in most cases, full implementations are optional and models as those presented here do their job.

References

Ashby, W.R., 1960. Design for a Brain: The Origin of Adaptive Behaviour, second ed. Chapman and Hall, London.

Beer, R.D., 1995. On the dynamics of small continuous-time recurrent neural networks. Adapt. Behav. 3, 471–511.

Beer, R.D., 2003. The dynamics of active categorical perception in an evolved model agent. Adapt. Behav. 11, 209–243.

Brooks, R.A., 1991. Intelligence without representation. Artif. Intell. 47, 139–159.

Di Paolo, E.A., 2000. Homeostatic adaptation to inversion of the visual field and other sensorimotor disruptions. In: Meyer, J.-A., Berthoz, A., Floreano, D., Roitblat, H., Wilson, S. (Eds.), From Animals to Animats 6: Proceedings of the Sixth International Conference on the Simulation of Adaptive Behavior. MIT Press, Cambridge, MA.

Di Paolo, E.A., 2003. Organismically-inspired robotics: homeostatic adaptation and natural teleology beyond the closed sensorimotor loop. In: Murase, K., Asakura, T. (Eds.), Dynamical Systems

Approach to Embodiment and Sociality. Advanced Knowledge International, Adelaide, pp. 19–42.
Di Paolo, E.A., 2005. Autopoiesis, adaptivity, teleology, agency. Phenom. Cogn. Sci. 4, 429–452.
Di Paolo, E.A., Harvey, I., 2003. Decisions and noise: the scope of evolutionary synthesis and dynamical analysis. Adapt. Behav. 11, 284–288.
Di Paolo, E.A., Rohde, M., De Jaegher, H., in press. Horizons for the enactive mind: values, social interaction and play. In: Gapenne, O., Stewart, J., Di Paolo, E.A. (Eds.), Enaction: Towards A New Paradigm in Cognitive Science. MIT Press, Cambridge, MA.
Goldstein, K., 1995/1934. The organism. Zone Books, New York.
Harvey, I.P.H., Cliff, D., Thompson, A., Jakobi, N., 1997. Evolutionary robotics: the Sussex approach. Robot. Auton. Syst. 20, 207–224.
Harvey, I., Di Paolo, E.A., Wood, R., Quinn, M., Tuci, E., 2005. Evolutionary robotics: a new scientific tool for studying cognition. Artif. Life 11, 79–98.
Iizuka, H., Di Paolo, E.A., in press. Towards Spinozist robotics: exploring the minimal dynamics of behavioural preference. Adapt. Behav.
Jonas, H., 1966. The Phenomenon of Life: Towards A Philosophical Biology. Northwestern University Press, Evanston, IL.
Kelso, J.A.S., 1995. Dynamic Patterns: The Self-organization of Brain and Behavior. MIT Press, Cambridge, MA.
Macinnes, I., Di Paolo, E.A., 2006. The advantages of evolving perceptual cues. Adapt. Behav. 14 (2), 147–156.
Moreno, A., Etxeberria, A., 2005. Agency in natural and artificial systems. Artif. Life 11, 161–176.
Nolfi, S., Floreano, D., 2000. Evolutionary Robotics. The Biology, Intelligence and Technology of Self-organizing Machines. MIT Press, Cambridge, MA.
Rohde, M., Di Paolo, E.A., 2006. 'Value signals': an exploration in evolutionary robotics. Cognitive science research paper 584, COGS, University of Sussex.
Rutkowska, J.C., 1997. What's value worth? Constraints on unsupervised behaviour acquisition. In: Husbands, P., Harvey, I. (Eds.), Proceedings of the Fourth European Conference on Artificial Life. MIT Press, Cambridge, MA, pp. 290–298.
Smithers, T., 1997. Autonomy in robots and other agents. Brain Cogn. 34, 88–106.
Thelen, E., Smith, L.B., 1994. A Dynamic Systems Approach to the Development of Cognition and Action. MIT Press, Cambridge, MA.
Thompson, E., 2007. Mind in Life: Biology, Phenomenology, and the Sciences of Mind. Harvard University Press, Cambridge, MA.
Tuci, E., Quinn, M., Harvey, I., 2003. An evolutionary ecological approach to the study of learning behavior using a robot based model. Adapt. Behav. 10, 201–222.
Varela, F.J., 1979. Principles of Biological Autonomy. Elsevier, North Holland, New York.
Varela, F.J., 1991. Organism: a meshwork of selfless selves. In: Tauber, A.I. (Ed.), Organism and the Origin of the Self. Kluwer Academic, Netherlands, pp. 79–107.
Varela, F.J., 1997. Patterns of life: intertwining identity and cognition. Brain Cogn. 34, 72–87.
Walter, W.G., 1950. An imitation of life. Scientific Am. 182 (5), 42–45.
Weber, A., Varela, F.J., 2002. Life after Kant: natural purposes and the autopoietic foundations of biological individuality. Phenom. Cogn. Sci. 1, 97–125.
Yamauchi, B., Beer, R.D., 1994. Sequential behavior and learning in evolved dynamical neural networks. Adapt. Behav. 2, 219–246.