Organismically-inspired robotics: homeostatic adaptation and teleology beyond the closed sensorimotor loop

By Ezequiel A. Di Paolo

School of Cognitive and Computing Sciences
University of Sussex, Brighton BN1 9QH, U.K.
Email: ezequiel@cogs.susx.ac.uk, *Fax: +44-1273-671320*

1. The problem of meaning in AI. Still with us?

In 1972, and later in 1979, at the peak of the golden era of Good Old Fashioned Artificial Intelligence (GOFAI), the voice of philosopher Hubert Dreyfus made itself heard as one of the few calls against the hubristic programme of modelling the human mind as a mechanism of symbolic information processing (Dreyfus, 1979). He did not criticise particular solutions to specific problems; instead his deep concern was with the very foundations of the programme. His critical stance was unusual, at least for most GOFAI practitioners, in that it did not rely on technical issues, but on a philosophical position emanating from phenomenology and existentialism, a fact contributing to his claims being largely ignored or dismissed for a long time by the AI community.

But, for the most part, he was eventually proven right. AI's over-reliance on world-modelling and planning went against the evidence provided by phenomenology of human activity as situated and with a clear and ever-present focus of practical concern – the body and not some algorithm is the originating locus of intelligent activity (if by intelligent we understand intentional, directed and flexible), and the world is not the sum total of all available facts, but the world-as-it-is-for-this-body. Such concerns were later vindicated by the Brooksian revolution in autonomous robotics with its foundations on embodiment, situatedness and de-centralised mechanisms (Brooks, 1991). Brooks' practical and methodological preoccupations – building robots largely based on biologically plausible principles and capable of acting in the real world – proved parallel, despite his claim that his approach was not "German philosophy", to issues raised by Dreyfus.

Putting robotics back as the acid test of AI, as oppossed to playing chess and proving theorems, is now often seen as a positive response to Dreyfus' point that AI was unable to capture true meaning by the summing of meaningless processes. This criticism was later devastatingly recast in Searle's Chinese Room argument (1980), and extended by Harnad's Symbol Grounding Problem (1990). Meaningful activity – that is, meaningful for the agent and not only for the designer – must obtain through sensorimotor grounding in the agent's world, and for this both a body and world are needed.

Following these developments, work in autonomous robotics and new AI since the 1990s rebelled against pure connectionism because of its lack of biological plausibility and also because most of connectionist research was carried out *in vacuo* – it was compellingly argued that neural network models as simple input/output processing units are meaningless for modelling the cognitive capabilities of insects, let alone humans, unless they are embedded in a *closed sensorimotor loop* of interaction with a world (Cliff, 1991). Objective meaning, that is meaningful internal states and states of the world, can only obtain in an embodied agent whose effector and sensor activities become coordinated

whilst performing the desired task. Neural network models of cognition can of course be abstract and simplified, but simplifying the sensorimotor loop out of which the contingencies and invariants necessary for stable perception and behaviour originate, we now believe, is akin to modelling supersonic flight and ignoring gravity.

Researchers working in new approaches to AI today (Pfeifer & Scheier, 1999; Harvey et al., 1997; Nolfi & Floreano, 2000; Brooks et al., 1999; Beer, 2000) are pretty well aware of these events that shaped the recent history of their field. Many of the problems outlined by Dreyfus became issues of (often practical) concern that had to be resolved typically with a turn towards aspects of real instances of cognition, what Varela called the re-enchantment of the concrete (Varela, 1995). This turn often became more compelling as we started to look at animal behaviour more closely. The exceptions provided by human intelligence, its ability for detachment and generality, have often been a stumbling block in the way to understanding the embodied and situated nature of cognition. The turn towards the concrete, in robotics, is mainly a turn towards the biological. And biological inspiration is now one of the proudest labels for the new robotics.

My purpose in this paper is to ask whether the slow unravelling of this story, the re-discovery through different avenues of what Dreyfus saw as the fundamental problems of AI already in the 1970s and which often has implied turning back to insights generated before the GOFAI age in cybernetics and holistic approaches to biology and psychology, has reached an end or whether something fundamental is still missing. I will argue for the second option. I think we have not yet seen the ending of this story, but that all the elements are in place at this very moment for moving on to the next chapter. As before, practical concerns will be strong driving motivations for the development of the necessary ideas. I will try to raise those concerns with the humble aim of putting them on the discussion table. I will not claim to solve anything although I will hint at possible solutions and discuss how good they seem to be at this stage.

In a nutshell, my claim is that autonomous robots still lack one fundamental property of what makes real animals cognitive agents. A property that motivated Dreyfus to argue against GOFAI. This property is that of *intentional agency*. Contemporary robots, I will argue, cannot be rightly seen as centres of concern, or put simply as *subjects*, the way that animals can. A robot failing in its performance does not show any signs of preoccupation – failure or success do not affect its structure in any dangerous way, nor does any form of concern accompany its actions simply because the desired goal is not *desired* by the robot but by the designer. Such robots can never be truly autonomous. In other words the presence of a closed sensorimotor loop *does not* fully solve the problem of meaning in AI. And this, as we shall see, is *not* merely a complexity issue (such is currently the widespread belief in the robotics community) but, as before, a question of not attempting to model flight without including gravity.

The problem, it is clear, will be to pin down such a general criticism to a concrete form, so that we can glimpse some technical response to the challenge. For this, I must be able to clarify what could one possibly mean with a statement such as "a robot has no intentions". I propose to do this by looking at the relation between intrinsic teleology and life as exposed in the philosophy of Hans Jonas (who is in fact representative of a larger movement of mainly Continental approaches to the philosophy of biology) and the convergence of these ideas with theories of self-organisation and autopoiesis; a convergence that has been described very recently (Weber & Varela, 2002). I will then ask the question of how close can we design a robot to not just to resemble but to *be* like an animal. This paper therefore falls within the class of recent efforts that argue for some sort of continuity between life and cognition (Maturana & Varela, 1980; Stewart, 1996; Wheeler, 1997).

The answer to these questions is definitely not of the additive kind. It will not be a matter of finding the missing ingredient so that it can be included in our robot design in a

box labelled Emotion or Motivation or Value System. Such attempts exist and generate results of interest which are perfectly valid in other contexts, such as modelling. But from the point of view of this discussion they miss the point, as I shall argue. The last part of this paper will deploy an alternative path based on Ashby's (1960) framework for adaptive behaviour applied to the process of habit formation, finishing with recent work on homeostatic adaptation (Di Paolo, 2000). Whether these solutions will prove useful is less important than raising awareness of the problem and generating debate.

2. The problem: In what ways are robots different from animals?

The label "biologically-inspired" has been applied to important advances in autonomous robotics, starting from Brooks' criticism of the sense-model-plan-act GOFAI approach to robot design as one that is unsupported by biological data (Brooks, 1991), moving on to the exploration of mechanisms, both neural and bodily, directly inspired on neuroscientific, physiological and ethological data to the effect of making robots more autonomous, more adaptable and more *animal-like*. An extremely fruitful way of working in synthetic approaches to robotics is the more or less systematic exploration of biological mechanisms not typically included in robot design. There's hardly a case where such an exploration, if done well, doesn't yield some interesting result.

In this respect it is possible to mention the work of Husbands and colleagues (Husbands et al., 1998) on GasNets, neural robot controllers where single nodes are capable of emitting and responding to diffusible neuromodulators, and the work of Floreano and Urzelai (2000) on the evolution of synaptic plasticity. Recent explorations have also turned to networks of spiking neurons (Floreano & Mattiussi, 2001) and spike-timing dependent plasticity (Di Paolo, 2003) by harnessing the power of evolutionary design to synthesize appropriate robot controllers which can later be the subject of analysis and may, eventually, feed information back to neuroscience by providing exemplars of whole-agent, closed-sensorimotor-loop control – something not typically explored in computational neuroscience. Biological inspiration has also influenced the design of robot body plans and sensorimotor arrays. Embodiment is often given a concrete, if perhaps sometimes limited, meaning by studying to what extent a body plan, a particular set of effectors or sensors, contributes towards the effectiveness of robot performance. Examples abound from simple, but powerful proofs of concept (e.g., Didabot, (Pfeifer & Scheier, 1999)), to complex insect-like and snake-like structures and humanoid robots. Often these designs are not static, but make use of passive dynamics principles and loose couplings through both the environment and other bodily structures to achieve robust and adaptable performance. In this respect, we can mention the whole sub-discipline of passive-dynamic walking (McGeer, 1990; Collins et al., 2001) and the work on joint control using bodily coupled neural oscillators (Williamson, 1998).

Further influences of a biologically-inspired frame of mind are sometimes subtler. Robot design is less seen as a task whereby a mechanism must control a body to achieve unconditional performance but, more often, the controller-body-niche coupling has become the object of design. A controller works with a specific body design as much as the body works with the controller. The robotic task is not expected to persist if the environmental coupling is changed radically, as much as an animal is not labelled as lacking in intelligence if its niche is fundamentally altered and its performance non-adaptive. This change of perspective is sometimes unspoken, a sort of growing consensus in how the aims of autonomous robotics should be approached. An excellent discussion and further review can be found in (Chiel & Beer, 1997; Beer et al., 1998).

Undoubtedly, there is still much to be done and discovered in this research approach to robotics. So, trying to point to its limitations could be seen at best as premature or, less charitably, simply as whining. I will try in the following to spell out in what important

ways biologically-inspired robotics is missing out on perhaps the most crucial aspect of living organisms, and what a radical change paying attention to this aspect could bring about. But in doing so, I am trying to add to what is currently being done, not to criticise it negatively, because the current approach has not yet reached a stage that needs such criticism. If it turned into a blind dogma, if it limited itself to merely copying just-another-unexplored-mechanism-found-in-animals, the same way that much of research in connectionism is limited to studying yet-another-variant-of-a-learning-algorithm, then perhaps a wake-up call would be necessary. So far, such has not been the case.

My contention, sketched here and further developed in the next sections, is that robot design may be getting important inspiration from the properties of biological neuronal mechanisms and from the dynamics of animal bodies, but it is getting little or no inspiration from the fact that the organisms that serve as models are *living systems* and that they structure their activities and their environment into a space of *meaning* which is defined as that which distinguishes between what is relevant and what is irrelevant to the organism's continuing mode of existence and ultimately survival. The crucial aspect that I would like to emphasise is that such Umwelt, in the words of von Uexküll (1934), is not divorced from the internal organisation of the organism. On the contrary, it is both generated by this organisation and causally connected to its conservation and subsistence, i.e., to what defines this organisation as that of a living being.

What an organism *is* and what it *does* are not properties external to each other that must be brought into coordination by some additional process (natural selection being the favourite candidate here). The organisation of a living system by itself creates an agency. By cleaving the material space into what the organism is and what is not, by forming such a concrete boundary, the living organisation generates and maintains a relation between the organism and its exterior. By being an organisation of continued material renovation, the relation generated with the outside is a relation of *need* and *satisfaction*. Need that originates in the thermodynamic constraints of a structure in constant material and energetic flow, satisfaction when those constraints are met.

In the particular case of animals, by putting a distance and a lapsus between the tensions of need and the consummation of satisfaction, the living organisation gives origin to a special relation with the outside, that of perception and action, which are charged with internal significance, and hence with grounded emotion. It is because of this that external events have concrete meaning for an organism – a meaning that is directly linked to whether its modes of being will continue or cease.

These statements need further unpacking, and I propose to do this in the next section, but we can already put some emphasis on the main issues: 1) an animal is a natural agent who generates its own boundaries and defines its world, a robot is an agent in virtue of external conventions of spatiotemporal and material continuity; 2) as a corollary, an animal does not simply *have* purposes, but *generates* them, a robot, alas, only has them, and it has them in a restricted sense solely in virtue of an externally imposed connection between its organisation and its environment. The relation of reciprocal causality that obtains in the animal, that between what it is and what it does and endures, appears in a broken version in the robot. Only the outward arm of the closed loop has occupied researchers: how to design a robot in order for it to do something. The way in which a robot's behaviour and the perturbations it must face may challenge and shape its organisation, that is, the inward arm of the causal loop that is the necessary condition for a natural agency to arise, has not been an object of design in itself (although some recent work is starting to move in this direction, see below).

Biological inspiration, at least up to this point, has thus been found to be of a limited scope – that of simulating and testing particular mechanisms found in animals. However, biologically-inspired robots are not necessarily *organismically-inspired*, i.e., inspiration has not come from the defining organisation and conditions of the living, or some of its

corollaries. In the following, I will try to expand on the seriousness of this criticism. At this stage it may seem a bit convoluted, or even pretentious, to point out that robots are not like living systems, but it will be possible to show that taking this point seriously will have significant effects of a practical nature, in particular having to do with robotic autonomy, adaptivity and intentionality, as well as effects on research and design methodologies.

3. Life, autonomy and identity

Before we formulate the question of whether it is possible to reproduce, at least in part, the fundamental organisation of living beings in an artefact, and what might the consequences of such an inclusion be, it is necessary to expend some effort in describing this organisation. This is a complex and rich subject which will by no means be exhausted in the next few paragraphs. I will only provide a sketch that will hopefully be complete enough for the purposes of the current discussion. The reader may wish to follow some of the references to find more extensive and profound treatments of the subject.

What is life? It is interesting how this question is more likely to be asked by a philosopher than a biologist. The historical reasons for this may turn out to be connected with the very answer to the question. Science seeks to explain phenomena in terms of underlying causal laws. If anything like circular causality or final causes are allowed, it is in the hygienized sense of "as-if" discourse. The term teleonomy (Nagel, 1977) is meant to precisely describe such an use. In this sense, we may speak of an organism's activities as directed towards an end, or of its morphology and physiology as serving a certain purpose, but in fact this is meant to capture regular correlations between an effect and its causes given a context of norms which is "tuned" – that is, brought into coordination – by some external process such as evolution and not internally generated by the organism itself. In this sense, an adaptive reaction in a predator anticipating a change in direction of the fleeing prey is something that has been acquired because it has been selected from a pool of random options as the one that better enhances the chances of survival. This etiological sense of functionality and purpose, see for instance (Millikan, 1984), has little to do with the question we are trying to answer. This sense of teleonomy, of "as-if" functionality, is equally applicable to non-living artefacts, as long as they can be shown to have the right historical events shaping their current structures and that their structures can be spoken about as serving functions whose origin are those same historical events. Such is the kind of functionality that may be found, for instance, in evolutionary robotics.

It makes sense also that we should apply a definition of life to situations where we know nothing about shaping historical events such as phylogeny. We should be able to recognize immediately other forms of life as living even though we may know nothing about their evolutionary history. In other words, a definition of life should be operational (Varela, 1979), i.e., based on what can be explained about a concrete instance without appealing to historical or contextual knowledge which by its very nature extends beyond the living system as it is given in front us into what has happened before of what may happen if we alter the current circumstances.

It is clear, however, that a biology based on a Newtonian model will be comfortable with the above sense of teleonomy and that issues that may challenge this situation are less likely to be openly discussed by biologists; for instance, speaking of organisms as natural and intrinsic purposes as Kant did in the second part of his *Critique of Judgement*. This is why philosophers, or philosophically minded scientists, have been more at ease with the question of life than many biologists. In the last century, an interesting convergence has occurred bringing together modern views of self-organisation and autopoiesis with existential and holistic takes on the question of life and behaviour. Scientists such as Plessner, Goldstein, Buytendijk, von Uexküll, Goodwin, Maturana, Varela, Rosen, and

Pattee have arrived, via diverse paths, to similar or close enough locations as philosophers such as Jonas, Merleau-Ponty and Straus. Such a convergence has been recently remarked by Weber and Varela (2002) in a discussion to which this paper owes much. The reader may also want to consult some of the main and secondary sources directly: (Maturana & Varela, 1980; Varela, 1979; Rosen, 1991; Goldstein, 1934; Goodwin & Webster, 1997; Uexküll, 1934; Straus, 1966; Jonas, 1966; Merleau-Ponty, 1963; Grene, 1968; Lenoir, 1982; Harrington, 1996). The most emblematic way of describing this convergence is in terms of an equal rejection of vitalism and other mystical and dualistic descriptions of life, and mechanicism, the Newtonian understanding of matter as inert and incapable of organisation unless this is provided by external forces.

Hans Jonas (1966) has examined the nature of organisms from an existential point of view. He puts the finger on metabolism as the distinguishing feature of life. A living organism does not posses the same kind of identity as a particle of matter. The latter is the 'same' entity in virtue of its spatiotemporal continuity. It is always 'this one' as opposed to 'that one' because of its material permanence. An organism, in contrast, stands out by the fact that it never actually coincides with its material constitution at a given instant. Due to its metabolism, its components are in constant flux. However, they maintain an organisation which assures its own durability in the face of randomising events that tend towards its dissolution. The organism has a formal and dynamic identity. It only coincides fully with its material constitution when it is dead.

Machines can also be seen as having a flux of matter and energy, typically in the form of inputs and outputs but, Jonas argues, the organism is unlike any machine in that the very nature of the flux is used to fabricate and maintain its own components. These are not externally given, nor have they any independent identity. If metabolism stops, the organism ceases to be (or adopts a non-living form with the potential of re-starting life at a later stage, such as a seed). If a machine stops, it simply stops, it does not turn into something else.

This can be re-stated as the fact that an organism's identity is given by its own functioning, or that the organism self-produces. At this point we may notice the strong parallel between this point of view and the one presented by the theory of autopoiesis. This framework developed by Humberto Maturana and Francisco Varela (1980, 1987) is based on the notion of autopoiesis, self-production, as the central property of living systems (as opposed to traditional "shopping list" definitions that merely state properties shared by many living systems). Accordingly,

> "An autopoietic system is organized (defined as unity) as a network of processes of production (synthesis and destruction) of components such that these components: (i) continuously regenerate and realize the network that produces them, and (ii) constitute the system as a distinguishable unity in the domain in which they exist", (Varela, 1997).

Formally stated, this embraces the definition proposed by Jonas and others such as Piaget (1967) back to Kant in the second part of the *Critique of Judgement*.

The more existential aspects of Jonas' biophilosophy should not occupy us here, but are worth mentioning because they are surprising. The first is that, admittedly, describing an organism in terms of its form, that is, as a centre of material and energetic flux, may after all be a matter of epistemological convenience. It is as describing a pattern of oscillating particles in a continuum as a wave. Once we know what each particle is doing, we know all about the wave, so the wave as an independent entity is just a useful, but fictitious, way of seeing the phenomenon of oscillating particles. Such, he argues, would be the way a mathematical mind would see life, as an epistemologically convenient embracing metaphor for particles that dance in a complex sort of wave. However, *we* can do better than that. We can ascertain beyond any shadow of a doubt that organisms

have an identity beyond the epistemological convenience of detached description. In living systems "nature springs an ontological surprise in which the world-accident of terrestrial conditions brings to light an entirely new possibility of being: systems of matter that are unities of a manifold, not in virtue of a synthesizing perception whose object they happen to be, nor by the mere concurrence of the forces that bind their parts together, but in virtue of themselves, for the sake of themselves, and continually sustained by themselves", (Jonas, 1966, p. 79). The way we know this for certain is simply that we *are* organisms. We know by the direct availability of our bodies and by our struggles that we are indeed one of these entities. We have, as Jonas puts it, inside knowledge – a knowledge that is not available to a disembodied and mathematical mind. This interesting existential turn in the argument, one that is difficult to argue with, is followed by an interesting corollary. If living systems are part of nature's ontology, and if they constantly depend on matter at a given time, but are not attached to a specific collection of particles through time, then their relation to matter, and so to the laws that govern matter, is one of *need* on the one hand and *freedom* on the other. Organisms are a wave of matter and energy, they are bound by the laws of physics but not fully determined by them as their destiny is not attached to any particular material configuration but they ride from one configuration to another. Jonas argues how this relation of needful freedom starts with metabolism but is later exploited and expanded by evolution in animals and eventually in humans. But this is beyond the scope of this paper.

More relevant for our purposes – determining what use we can make of these ideas for designing organismically-inspired robots – is the following question: Why is the living organisation so special and difficult to describe? Because it refers at the same time to the processes of constitution of the organism, processes that define what the organism is, and to how those processes are generative of the organism as an agency, i.e., what the organism does. The two are complementary aspects of the same entity. A typical systemic description starts with a system which is well-defined in terms of components and their relations. A living system is such only in virtue of the constant threat of becoming a different system or even a dead one. We may observe some stability in the organisation of an organism and be happy to constrain a model spatially to unchanging fragments of the organism or temporally by assuming that the organism will not change significantly in a given period. But we cannot truly model what is proper to the living organisation unless we leave room in our descriptions for the fact that this system could, at any given time, become something quite different. This threat, even if kept at bay by the adaptivity of the living organisation, is (negatively put) part of the defining property of life. We cannot understand life conceptually otherwise. Life, in short, is defined in terms of death and the threat of change.

Allowing the possibility of ongoing shifts in the identity of the system in question, makes it quite difficult for scientists to model life (or closer to our concerns here, for a roboticist to create an artificial living robot). This is because all models must be grounded on a concrete description of the system being modelled. In other words, they rely on an identity of components based on continuity, and not on the self-generated identity of the organism. Living systems may change their structures as long as the autopoietic core is maintained. This means that the description of the system is never fixed but subject to constant re-structuring. Components may come and go, and their relations change. Eventually even the autopoietic organisational core will be lost (when the system dies). This means that a successful model of a living system should involve a description of the system while the description itself is contingent to what the system does! Ultimately, when death arrives, the model should be not of a "system" at all, but of aggregates of inanimate components. Some people think such a modelling feat is impossible, I say that the very least is extremely difficult to conceive within the current modelling paradigms. Simulation modelling, with its potential for capturing different

levels of organisation, may be the answer (Di Paolo et al., 2000). This is because it is possible to model fixed properties in components that act as constituents of higher level systems. Some existing simulation models are concerned with these questions (Varela et al., 1974; Fontana & Buss, 1996; McMullin & Varela, 1997; Rasmussen et al., 2001) but so far nothing resembling a full model of a living organisation has been produced.

4. Survival: "the mother-value of all values"

What an organism does (both as complex metabolic system and as a natural agency in its world) is to actively seek its own continuation. Those aspects of its interactions that contribute to this natural purpose are seen as intrinsically *good* according to this self-generated norm. And those aspects that challenge this end, are intrinsically *bad*. Thus, a purely dynamical principle of self-continuation engenders an intrinsic teleology.

This perspective has introduced an important concept that has been very useful, if not often stated explicitly, in the study of adaptive behaviour: the viewpoint of system viability as the generating value for adaptation. According to this concept, behaviour can be classified as adapt*ed* if it contributes to the continued viability of a system, and as adapt*ive* if it restores the necessary stability when this viability is challenged. Such is the basis of the framework for adaptive behaviour developed by Ashby in the 1940s.

Ashby (1960) saw the survival condition as being represented by the state of what he called essential variables. Examples of these for mammals are sugar concentration in the blood, body temperature, etc. If these variables go out of bounds the system's survival is challenged. If the adaptive response cannot be fully generated internally (by means of regulative processes) and must be accomplished via interaction with the environment, the resulting behaviour is seen as adaptive. Now there are two possibilities: either 1) the mechanisms that generate such adaptive behaviour are already in place, e.g., the animal knows how to search for food when hungry, or 2) they are not, in which case they must be generated somehow if they organism is to subsist.

Ashby proposed a general explanation for the second case. He linked the condition of the essential variables going out of bounds with the mechanisms for testing new behavioural parameters. Once a set of parameters is found so that stability is restored to the essential variables, the search is over, and the organism has adapted. Systems capable of such re-organisation are called ultrastable. In figure 4 we reproduce Ashby's explanation. The dashed line represents the organism. Within it, R is its behaviour-generating subsystem, P represents those parameters controlling R (i.e., changes in P will affect R and consequently behaviour), and the meter represents the essential variables. The organism is in a two way interaction with its environment (represented by the arrows connecting Env and R). The environment may also affect the essential variables (arrow to meter) by posing challenges to the organism (a poisonous food, a predator, etc.) and these in turn affect the parameters controlling behaviour. Ashby's interesting choice for how the essential variables affect P is to use random step functions. If a new dynamics leading to a stabilization of the essential variables exists, then it will be found eventually by this process of random search and the system will be ultrastable.

The choice of random step-functions is conceptually interesting† not because we can argue that random search is indeed the case in real organisms, but as a proof that dumb mechanisms can yield adaptive responses which from the point of view of an external observer may look quite clever. Ashby was one of the first to challenge the viewpoint – that started with that battlehorse of the cybernetic era, McCulloch and Pitts' network of binary gates, and its heavy inspiration in the logical revolution of the 1930s and continues to this day – that intelligent performance needs to be the result of equally

† These mechanisms were built into an actual ultrastable electromechanical device called the Homeostat, (Ashby, 1960).

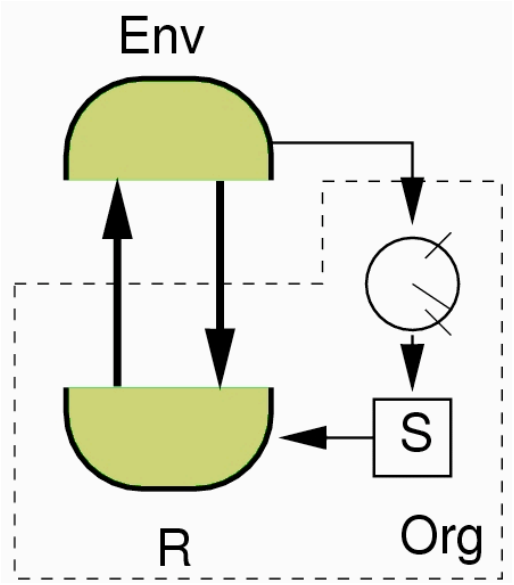


Figure 1. Ashby's concept of an ultrastable system. The organism is represented by the dashed line. R represents the behaviour-generating sub-system of the organism which is in a closed sensorimotor loop with the environment (Env), P represented those parameters that affect R, and the meter represent the organism's essential variables which can be directly affected by the environment and in turn affect P by introducing random step-changes in the parameter values. If these changes produce a new dynamics capable of restoring the essential variables to their equilibrium through the newly acquired behaviours, the organism will have adapted to the environmental challenge.

intelligent mechanisms. It needs not be like this, and it is quite likely that it is not in general‡, and such is an important lesson for anyone interested in designing robots.

Another important lesson we can draw from Ashby's framework before we turn to a discussion of the limitations of the viability perspective, is that a closed sensorimotor loop is *not* enough for adaptation, but that at least a double feedback structure is needed. Notice in figure 4 that R is already in a closed sensorimotor loop with Env, but that this by itself does not guarantee the generation of the right response if a new environmental disturbance appears. This in other words means that the significance of an event or an internal state, whether it challenges or not the continued existence of the system, is not fully defined by a closed sensorimotor loop as is often implied by roboticists. A closed loop would seem to be necessary for grounding meaning, but not sufficient, and this is one of the topics we will develop in the next section.

5. Can artefacts have natural purposes? The ecological identity of habits and why emotions don't come in boxes

In contrast to organisms, the robots of today are "agents" in an impoverished sense of the word. Even when embedded in a sensorimotor loop of situated interaction their behaviour can be fully described simply as movement as opposed to action. Their "world" is only the set of external variables that may affect such moving trajectories. As worlds go, a

‡ A nice example is afforded by insect navigation using path integration and landmark recognition. Feats of navigation that seem to require a sophisticated mechanism capable of managing and retaining topographical information, for instance in the Tunisian desert ant *Cataglyphis*, can in fact be explained by a combination of distance integration, use of celestial compass and recognition of proximal landmarks, (Wehner et al., 1996).

robot's is quite devoid of significance in that there is no sense other than the figurative in which we can say that a robot *cares* about what it is doing. A robot may show cleverness of design and some degree of adaptivity, but this by itself still puts it not in the same class as animals but closer to the thermostat. What is missing?

Many roboticists either know or intuit this problem. The assumed answer to this question is that robots lack the degree of complexity of real animals, and that what is necessary is to keep climbing the complexity ladder. I believe this is true. I also believe it is the wrong answer. It is wrong in the sense that it is not the most practical answer, the answer that will lead us to question in detail what is at the source of the problem. Indeed, the solution or solutions to this problem will require more complex designs for robot bodies and controllers, more complex tasks, and more complex methods of synthesis. But seeking such complexity blindly, by typically restricting the search to achieving more complex behaviours, does not accomplish much.

To try to spell out the question that I believe is blocking the development of organismic-like robots it will be useful to pose a much more concrete question: Why is light meaningless to a Braitenberg vehicle†? At first sight, this is a strange question. After all, what else is there in the world of a Braitenberg vehicle other than a smooth surface and sources of light? However, unlike the situation we have described above for organisms, a Braitenberg vehicle has no stake in the continued performance of its behaviour. Swapping the position of the sensors will result in an immediate change of behaviour which will not destroy the robot, nor will it put it in any danger of ceasing to be what it is. Remove one motor and the robot will rotate on the spot and will be equally happy with this behaviour as it was before when it was performing phototaxis. This situation is no different for more complex robots.

In this context, it is interesting to notice that Braitenberg described the behaviour of his vehicles as invested with emotional interest with inverted commas. He spoke of robots as showing "love" and "fear". I think this charming description is challenging in two possible senses – one of them is justified but not the other – and these should be kept distinct. The first sense is that complex behaviour needs not follow from equally complex mechanisms (recall Ashby's random step-functions). This has been taken as a lesson not to be forgotten by new AI and the autonomous robotics movement. The details of the body and the environmental coupling are sources of complexity and subtlety in themselves. The controller may be as simple as a wire going from sensors to effectors and yet behaviour may be interesting enough to deserve descriptions typically reserved for higher animals.

However, there is another way of interpreting Braitenberg ascription of emotion to his robots, one that is more literal and less justified. This sense contends that even though these robots display "love" or "fear" only with inverted commas the way towards real affective behaviour is simply the path of increasing complexity, and that if a robot is able to convincingly behave as if it had emotion, then this will be a sufficient criterion to assert that in fact it does. In other words, inverted commas can be erased if the behaviour is convincing and complex enough.

This interpretation has guided much of the work on robot emotion which is mostly concerned with functional and imitative considerations. Emotion is linked to behaviour by functional links, i.e., an affective state is brought about by behavioural circumstances, like the perception of danger, and it modulates the generation of behaviour accordingly, for instance by entering into a state of alarm. It is also important that these pseudo-

† Braitenberg vehicles, very much used in robotics as conceptual and practical tools for design, are the simplest wheeled robots one can conceive. Light sensors are wired directly to motors either on the same or opposite sides of the body and either with negative or positive transfer functions that regulate motor activity, thus resulting in robots that modulate their behaviour with respect to light sources either by avoiding or following them, (Braitenberg, 1984).

meaningful spaces with their dynamic rules be convincing, such as in the work on Kismet at MIT (Breazeal, 2001) and others. But these emotional spaces and their links to behaviour are externally defined. Emotion is simply simulated. The emotional links between the agent's behaviour and the internal mechanisms are arbitrary, in the sense that even though it is possible for Kismet to express fear or embarrassment if the human interlocutor speaks in an angry tone, it is equally easy to make the robot look happy under the same circumstances. A real animal, however, can be trained to do lots of things, but never to treat a punishment as a reward. The link between the external situation, the internal dynamics and the overall affective state may be contingent (different species will look on the same piece of fungus with desire or repugnance), but never arbitrary as it is defined by the continuous existence and renovation of the autopoietic organisation and subordinated to it and to the network of interacting tensions and satisfactions it creates in a behaving animal. To believe otherwise, to think that adding a box labelled "Emotion" to the robot's controller is enough for a robot to *have* emotions, is to commit a fallacy of misplaced concreteness‡.

But surely we have not misused the word adaptation to describe the behaviour of autonomous robots. Are they not able after all to regulate their actions and guide them according to contingencies so as to successfully achieve a goal? Are not Braitenberg vehicles, radical sensor and motor perturbations apart, capable of reliably approaching a source of light while coping with deviations in their trajectories? Why is this not enough to grant them some degree of teleology?

The answer again is that closed sensorimotor loops fall short of providing true intentionality to a robot's behaviour. The "solution" to the problem of purposes in nature proposed in the seminal paper by Rosenblueth, Wiener and Bigelow (1943), that of feedback loops providing the sufficient mechanisms for goal-seeking behaviour, became acceptable for roboticists and has never been questioned since†. It has, on the contrary, often been forgotten and researchers have rightly advocated for its return (Cliff, 1991; Pfeifer & Scheier, 1999). The importance of breaking with a linear way of thinking about the generation of behaviour – one that goes in an a open arc from sensations to motor activity – cannot be underestimated and in the current context in robotics and cognitive science, must still be firmly stressed. This was the concern of Dewey's (1896) criticism of the reflex-arc concept in psychology and Merleau-Ponty's emphasis on the motor basis of perception (Merleau-Ponty, 1963). Both these philosophers, however, went further than being content with closed sensorimotor loops in the sense that for both perception and action are accompanied by intention and extend into the the agent's situation and history. The actions of an intentional agent are charged with meaning.

To be fair, it is important to distinguish this natural sense of meaning from meaning

‡ A similar criticism applies to much of the interesting work done using value systems to regulate plastic change (Reeke Jr. et al., 1990; Salomon, 1998; Pfeifer & Scheier, 1999). Here, the system does not generate its own values but these are externally defined. The justification is often provided in evolutionary terms given that such regulatory structures would have evolved to guide learning towards survival-enhancing options. The evolutionary factor is not denied here, but again, it is insufficient for accounting for how an actual instance of a living system generates those values in the here and now, that is, for explaining intrinsic teleology. See also (Rutkowska, 1997).

† Jonas (1966) dedicates a chapter to strongly criticise the cybernetic idea of feedback loops as generators of natural purposes. Indeed, such systems do not differ from a mechanical system reaching its point of equilibrium. They may indeed provides us with goal-seeking mechanisms, but there is no sense in which they define and generate their own goals. Neither can these mechanisms generate behaviour that we would qualify as *unsuccessful* goal-seeking. A feedback system either causes a goal state to happen or it does not. It is only possible to ascribe goals to the system if these are reached. Any talk of the system trying to achieve a goal state but failing is metaphorical. This criticism parallels others made by Taylor (1964) and Wright (1976). Not knowing how the AI and connectionist movements would tend to ignore the importance of sensorimotor loops, Jonas was perhaps too severe with respect to the significance of the cybernetic argument. His point, however, remains. Closed feedback loops are certainly necessary for purposeful behaviour, but not in themselves sufficient.

as objective grounding in sensorimotor capabilities. The latter is a correlation between sensors and effectors that is the internal counterpart of a correlation between agent an environment so that the appropriate behaviour is enacted in a given situation. This is the sense that Cliff (1991) was probably concerned with in his arguments against disembodied connectionism. In the presence of a sensorimotor loop neural patterns of activations are not arbitrary with respect to their sensorimotor consequences and the future history of activation. The output which is reflected in motor activity modifies the input in a coordinated and contingent manner. The high activation of the proximity sensor followed by the appropriate motor action will turn into low activation. This correlation is what is meant by sensorimotor coordination and it may take much more complex forms, (Scheier & Pfeifer, 1995).

But there is another sense of meaning, one that is closer to the intrinsic teleology of living systems. This is the sense in which a situation is meaningful in terms of its consequences for the conservation of a way of life. In the living system this is ultimately the conservation of its autopoietic organisation, its own survival and viability. But this answer leads us to an issue of practical importance. How can we invest artefacts with a similar sense of meaning? Do we need to go all they way down to metabolism and self-production or is another solution possible?

According to the viability perspective sketched above and advocated not only by Ashby but implicated as well in the biophilosophy of Jonas and autopoietic theory, survival (in the sense of continuity of metabolism) would seem to answer for the origin of natural norms and would indeed count as the mother-value of all values (Weber & Varela, 2002). However, this proposal is not without shortcomings in explaining actual instances of behaviour. Behaviour is often underdetermined by the condition of continued viability. Finding food is necessary for survival, but often there are different sources of food and different ways of obtaining it. In spite of this, organisms will typically choose one behaviour out of all the possible viable ones. Understanding the reasons behind this fact may well provide us with the tools for building intentional robots without them needing to be alive in the autopoietic sense. Survival may be the mother-value of all values but the daughter-values have enough independence of their own. The key is that robots need not be required to conserve a *life* they lack but, as hinted above, *a way of life*. In order to understand this we must turn to the nature of preferred behaviours and the process of habit formation.

For seeing how adaptive behaviour is underdetermined by survival let us pay attention to one of Kohler's (1964) experiments on visual distortion (see also next section). The experiment involved a subject wearing special coloured goggles that were blue for one half of the visual field and yellow for the other half. After a few days the subject recovered normal colour vision and on removing the goggles he experienced an opposite change in hue depending of whether he looked right or left by moving his eyes with respect to his head (but not by keeping them fixed and moving his head). This is a fascinating result but only one aspect concerns us here: the visual perturbation introduced during the experiment cannot be said to have challenged the survival of the subject in a direct manner. We cannot account for the situation in the same sense as Ashby's essential variables going out of bounds. (Some variables may have gone out of bounds but they were not really *essential* in that we can imagine that life could have continued without the subject experiencing any adaptation, at most it would have been strange for him, but not physiologically dangerous). But then what norm could guide adaptation in such cases if not survival?

The answer lies in the objective ecological structures that underly the enactment of behaviour. We can afford to think in an Ashbyan style for explaining how behaviours (in the general sense of actions and perceptions combined) tend to get increasingly attuned to the regularities of the body and its surrounding so as to achieve what Merleau-Ponty

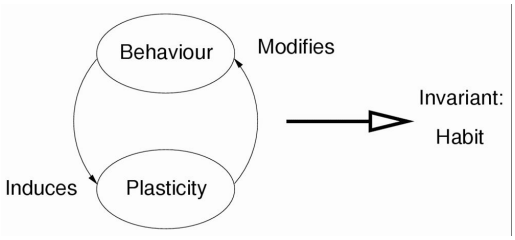


Figure 2. Habit as the invariant obtaining from two-way links between plasticity and behaviour under recurrent situations. It is implicit in this diagram that the resulting habit does not run counter the viability condition.

denominates *maximal grip*. If we assume that the potential for plastic change is ever present, and that plasticity affects the generation of behaviour which in turn affects plasticity in an activity-dependent manner and through recurrent engagements in similar situations, then two scenarios are possible: 1) the double link between plasticity and behaviour never reaches a stable dynamical regime, in which case it will inevitably run into the mother-value of all values and so be subject to an adapt-or-die situation; or 2) it may reach a stable regime and this may be unadaptive (with the same result as above), or adaptive in which case it will be *conserved* in preference to other *equally viable* behaviours. This invariant result (not a static structure but a dynamic one predisposed towards its own continuation, similar to a melody where the parts coordinate with one another and anticipate what is going to happen) is called a *habit*, (see figure 5). Notice that such a general explanation suffices to solve the problem of preferred forms of motor coordinations and the origin of motor synergies given the redundancy of the animal musculoskeletal system, also known as Bernstein's problem (Bernstein, 1967). And it goes beyond this by equally accounting for constancies in perception, albeit in an abstract manner†.

Now, the relevance of such structures has been identified and defended before, e.g., (James, 1890; Dewey, 1922, 1929; Goldstein, 1934). Piaget (1948, 1967) has provided a detailed account of the processes of equilibration leading to their stabilization and incorporation into an existing repertoire of behaviours. Habits, as self-sustaining dynamic structures, underly the generation of behaviour and so it is *them* that are challenged when behaviour is perturbed. An interesting hypothesis is that often when adaptation occurs in the animal world this is not because organismic survival is challenged directly but because the circular process generating a habit is. This may help us understand everyday instances of adaptation, such as re-learning of normal sensorimotor patterns after an injury or body reconfiguration, or adaptation to radical changes in our surroundings, such as moving home or country. We may invest our robots not with *life*, but with the mechanisms for acquiring *a way of life*, that is, with habits. This may be enough for them to generate a natural intentionality, not based now on metabolism, but on the conservation of 'one'

† It is interesting that Ashby himself was not very strict with his concept of essential variables. These are in principle variables *of* the system that adapts (e.g., physiological variables in an organism) and not external to it. He, however, used the term more loosely (Ashby, 1960, S 17/4) which naturally leads to an extension of his framework to the process of habit formation. Goldstein (1934) himself provides an account of learning to ride a bycicle by quasi-random trial and error that very much resembles the above proposal. In proposing this view on habit formation, we are simply saying that the underlying processes are autonomous in the sense of organisational closure proposed by Varela (1979). Biological autonomy of this kind is a generalisation of the idea of autopoiesis to self-sustaining networks of processes which may be generic, and not just of production and transformation of material components as in the case of metabolism. Our claim may be reformulated as stating that habitual patterns of behaviour have a natural identity given by the autonomy of their formation, and so they can, as metabolism, serve for grounding teleology and explaining adaptation to disturbances that do not directly challenge the survival of the organism.

way of life as opposed to 'another one'†. How this sort of intentionality relates to the intrinsic generation of intentions by a metabolizing system remains an open issue that need further development.

If such a proposal is adopted, the problem space will have changed in interesting ways. Because, habits may indeed die out without implying the death of the system. They may drive the system to situations that are contrary to its own survival or well-being (think for instance of addictive or obsessive behaviour). The interaction and commerce between these structures of behaviour, and not this or that particular performance, would become the object of robotic design, and the conservation of an organised meshwork of habits, the basis on which to ground artificial intentionality.

One trivial, yet important, corollary of this proposal is that plasticity is necessary for intentional agency. Change must be a possibility for a pattern of activity to try to avoid it, i.e., conservation only makes sense in the face of change. This may turn out to be a central pragmatic principle for designing intentional robots, they must be constantly challenged, in their bodies and controllers, as well as their patterns of sensorimotor coordination, so that they dynamically can resist those challenges, not only by being robust but also by adapting to them. Adequate research methodologies for this challenge are barely starting to be put to the test.

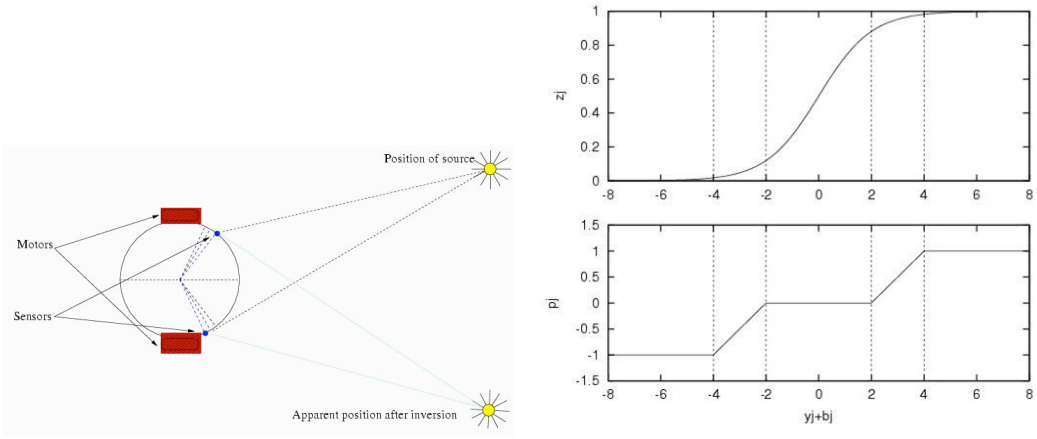
6. Homeostatic adaptation

The artificial evolution of homeostatic controllers for robots (Di Paolo, 2000) was originally inspired by the work of James G. Taylor (1962) who presented a theory based on Ashby's concept of ultrastability to account for the fascinating results obtained in experiments on adaptation to distortions of the visual field. Taylor worked in Innsbruck with Erismann and Kohler on adaptation to different sorts of prismatic perturbations including full up/down and left/right inversion. These experiments involved subjects wearing distorting goggles for a few weeks and are very well described in (Kohler, 1964). They follow in style similar experiments by Stratton in the 1890s and Ewert in the 1930s. The general pattern in all these experiments is that the subject starts by being extremely disoriented at the beginning of the experiment, everyday behaviours such as walking in a straight line become very difficult. Over the first few days, the subject slowly experiences improvement involving conscious effort initially, but becoming less and less conscious with time. Adaptation occurs eventually in those patterns of behaviour that the subject engages in, which may include walking in busy roads, running, riding a bicycle and even skiing. However, transfer of adaptation from one pattern to another is not guaranteed – in other words, habits must adapt more or less independently. Towards the final stages of the experiment, the subject has become used to the distorted visual field, and even reports that "it looks normal". The extent of this perceptual adaptation has been debated. For instance, it never seems to be complete, but partial. Kohler reports on a subject who, wearing left/right distorting goggles, saw cars coming up the "right" side of the road, but the letters on their plates still looked inverted which indicates that visual perception is far from uniform but formed also out of different habits.

The experiments below, carried out on a simple simulated robot platform, pale in comparison with the richness of these results. However, they should be seen as proof of concept; in particular, the concept of ultrastability which provides a way in for addressing some of the questions raised in this paper.

The initial model involved evolving the neural controller for a simple wheeled robot whose task is to perform phototaxis on a series of light sources randomly placed in the

† I'm leaving open the issue of how such robots might eventually be applied to useful purposes, which is not trivial. A truly autonomous robot may need to *convinced* to undertake a task that might be perceived as dangerous for its way of life and viability!



(a) (b)

Figure 3. Experimental scheme: (a) robot body and effect of sensor swapping; (b) facilitation of local plasticity as a function of 'cell potential' (y). Top: neuron activity ('firing rate'). Bottom: strength and sign of local plastic facilitation.

environment. One source is presented at a time, the robot must approach it, and after a while the source is put out and another one appears in the distance. The robot uses two motors and two light sensors (figure 3a). In addition, the robot controller is also evolved so that on average each neuron will maintain a firing rate that is neither too high or too low. Whenever a neuron fires above or below a given threshold, it activates genetically specified rules of synaptic change on its incoming connections. This introduces a potentially ultrastable element in each neuron, á la Ashby. This choice is not without biological justification – theoretical arguments and empirical findings confirm that cortical neurons tend to regulate their firing rates by similar mechanisms (Horn et al., 1998; Turrigiano, 1999). Figure 3b shows the sigmoid activation function for each neuron and the corresponding level of plastic facilitation p which can be positive or negative in this initial model. A central region where $p = 0$ marks the area of stability where no plastic change occurs. Fitness points are given for the time neurons spend within this region.

Once the robots have been evolved for both phototaxis and homeostasis we study the robot under the condition of visual inversion by swapping the sensors left and right (figure 3a), and observe whether the internal changes driven by loss of homeostasis are able to induce the recovery of behavioural function (in this case some form of phototaxis).

A fully connected, 8-neuron, dynamic neural network is used as the robot's controller. All neurons are governed by:

$$\tau_i \dot{y}_i = -y_i + \sum_j w_{ji} z_j + I_i; \quad z_j = \frac{1}{1 + \exp[-(y_j + b_j)]}$$

where, using terms derived from an analogy with real neurons, y_i represents the cell potential, τ_i the decay constant, b_i the bias, z_i the firing rate, w_{ij} the strength of synaptic connection from node i to node j, and I_i the degree of sensory perturbation on sensory nodes (modelled here as an incoming current) which is always 0 for the other neurons. There is one sensory neuron for each sensor and one effector neuron for controlling the

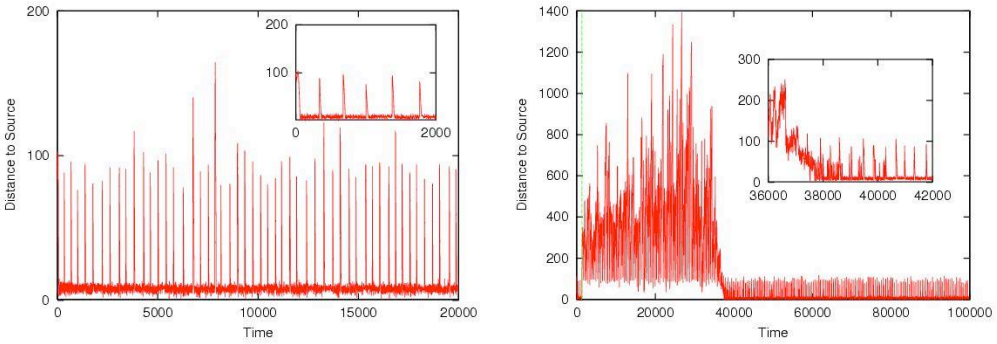


Figure 4. Robot's performance during the presentation of a long series of light sources: (a) distance to source under normal conditions (50 sources); (b) distance to source after visual inversion (250 sources) showing adaptation. Insets show details of the same plots.

activity of each motor. The possible rules of synaptic change are:

$$\begin{aligned} R0 : \Delta w_{ij} &= \delta\, \eta_{ij}\, p_j\, z_i z_j, \\ R1 : \Delta w_{ij} &= \delta\, \eta_{ij}\, p_j\, (z_i - z^o_{ij}) z_j, \\ R2 : \Delta w_{ij} &= \delta\, \eta_{ij}\, p_j\, z_i (z_j - z^o_{ij}), \\ R3 : \Delta w_{ij} &= 0, \end{aligned}$$

where Δw_{ij} is the change per unit of time to w_{ij}, δ is a linear damping factor that constrains change within allowed weight values, and p_j is the degree of local plastic facilitation, explained above (figure 3b). All rules and neural parameters are evolved using a typical genetic algorithm. See (Di Paolo, 2000) for more details.

Robots were first tested for long-term stability. During evolution they were evaluated on a series of 6 light sources, which is no guarantee that they will perform phototaxis if presented with, say, a sequence of 200 light sources. Of the runs that were long-term stable, it was found that about half of them were also ultrastable by adapting to visual inversion. Figure 4 shows one robot's performance with and without visual inversion. The plots show the distance to the active light source. During normal phototaxis this distance decreases rapidly and the robot remains close to the light until a new source appears in the distance, thus producing a series of peaks over time. Visual inversion produces the initial effect that the robot moves *away* from the light (as effectively right and left directions have been swapped), and keeps moving away until eventually adaptation ensues, and the robot begins to perform normal phototaxis again. Adaptation occurs because visual inversion has driven neurons close to a region of instability and so this eventually affects the network structure by the facilitation of plastic change. Because we have demanded of evolution to produce two fitness requirements – internal stability and phototaxis – in the same controller, it is likely that both conditions will require one another. In this case, regaining internal stability also means performing phototaxis again. Similar adaptation occurred in the presence of other perturbations.

The external behaviour of the robot resembles that of an agent whose goal is internally generated. However, we may identify a series of problems with the above scheme.

1. Variability: the rules of plastic change tend to produce inherently stable weight dynamics. This contrasts sharply with the pure random change scenario advocated by Ashby. In particular, three main factors have been identified: a) the lack of a stochastic element in weight change, b) the use of positional damping factors which tend to produce bimodal weight distributions, where a weight will be most likely to

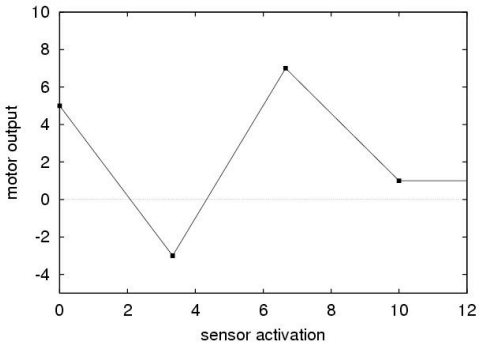


Figure 5. Example of mapping between one sensor and one motor, squares indicate the parameters subject to random change (applied only to y-coordinates in the results shown). Each motor is controlled by a piecewise linear surface which is a function of sensor activation.

be either at the maximum or minimum of the range, and c) the use of genetically specified initial weight values, and initial directions of weight change, thus providing an innate bias for determining weight values and removing the opportunity for activity-dependent change.

2. Task: The robot task is indeed quite a simplified version of the experiments on visual inversion. Although there are many different strategies for approaching a source of light, the possible behavioural classes are only three: a) approach, b) avoid, c) ignore the light. A richer scenario would include a better model of vision and more dimensions to behaviour, for instance using receptive fields and an arm model that permits the study of alternative sensorimotor habits and their recovery under more subtle forms of sensor distortion.

3. Performance and internal stability are disjoint: Perhaps most important in the context of this paper is the fact that the robot is not designed to have a two way causal link between its conditions of subsistence (internal stability) and its condition of performance (stability of the behavioural pattern). This is not something that was required in the Ashbyan framework, but the use of random search allowed for eventual adaptation whenever possible. In real organisms, however, the situation is different as we have seen. The robot is asked to meet two different requirements. Evolution may come up with two possible classes of solutions to this problem: a) internal and behavioural stability require one another, b) internal and behavioural stability simply do not interfere with each other. In the first case, we shall observe instances of homeostatic adaptation, in the second we shall not, as robots are capable of regaining internal stability without altering the perturbed behaviour. It would be much better if we could design a scenario where a) always holds.

The following experiment is an attempt to deal with problems 1 and 3 in the list. Problem 2 has not been addressed yet. In fact, the scenario is made even simpler by working with a Braitenberg architecture with plastic transfer functions between sensors and motors. In order to build in a two-way link between performance and internal dynamics we can imagine a light-powered robot with an internal battery that discharges with time and can be re-charged by approaching a source of light. The homeostatic variable becomes the level of the battery, and what is regulated by it is the mapping from sensors to motors. Thus the only stable condition, both internally and in terms of behaviour, is to perform phototaxis.

Both sensors connect to the two motors. Each mapping of motor vs. sensor activity is built by a piece-wise linear function fully specified by 4 points in the plane. The two points corresponding to the extremes have coordinates $(0, y_1)$ and $(10, y_4)$, and the two other points in the middle have coordinates (x_2, y_2) and (x_3, y_3), thus defining a mapping like the one shown in figure 6. Each motor averages the signal it receives from each sensor. The battery level E is governed by:

$$\tau_E \frac{dE}{dt} = -E + 10\frac{S_1 + S_2}{2}$$

where S_1 and S_2 are the activation level of the sensors and $\tau_E = 2500$ is the decay constant of the battery. A light source is presented on average for a time of 100, so that the timescale of battery recharge is roughly the same as 2.5 light source presentations and the timescale of discharge is 10 times as long. Sensor values are clipped to a maximum of 10.

The robot is initialised with random mappings. As long $E_{min} < E < E_{max}$ no change is introduced to the configuration of the controller. Outside these bounds, random change is applied to the parameters defining the four mappings. Two conditions have been studied with similar results: soft and hard boundaries. In the latter case, as soon as E goes out of bounds random change is applied by adding at each time step ($\Delta t = 0.2$) to each parameter value a normally distributed random variable with zero mean and deviation $\sigma \in [0.0005, 0.001]$. All the parameters are encoded by a number between 0 and 1 and then appropriately scaled. Reflexion is applied if a parameter falls outside this range. If boundaries are soft, the deviation is multiplied by a factor going from 0 at the boundary to 1 at a value outside the boundary – the corresponding intervals are (E_1, E_{min}) for the lower bound and (E_{max}, E_2) for the upper bound. Values have been successfully tried in the following ranges: $E_1 \in [0.1, 0.5]$, $E_{min} \in [2, 5]$, $E_{max} \in [10, 80]$, and $E_2 = E_{max} + 10$ (the high values in the last two cases correspond to effectively not having an upper boundary). The results below use soft boundaries although using hard boundaries does not introduce any qualitative difference.

As all parameters are pre-defined or defined on the run, there is no need to optimise the system using a GA (although optimisation could certainly be applied to some of the above parameters). Figure 6 shows the battery level of a long run consisting of 20,000 light sources. The horizontal line indicates the lower boundary of the essential variable. The higher boundary is set to 20. In the middle of the run (vertical line) the sensor positions are inverted, the robot loses the acquired behaviour, but regains it afterwards. The first half of the run shows an incremental level of adaptation with the essential variable drifting "inwards" into the viable zone. This has been observed repeatedly (but not always) and can be explained in Ashbyan terms (Ashby, 1960, S 14/3) as the controller jumping from one pocket of stability into another and remaining for longer in those pockets that afford increasingly greater robustness against the random elements of the environment (sensor and motor noise, random positioning of new lights, etc.). See also (Ashby, 1947). Figure 6 shows a detail of the distance to source and the value of E while the robot is phototactic. Notice that not all the light sources are approached, but that the battery level is high enough to afford missing the odd source.

The robot will also reproduce Taylor's (1962) experiment where he wore left/right inverting goggles in the morning and removed them in the afternoon for several days. He eventually became adapted to both conditions. (And able to put on and remove the goggles while riding a bicycle!) If every new source of light the sensor positions are swapped, the robot eventually finds an internal configuration able to cope with both conditions (figure 8).

This scheme gets around problems 1 and 3 and is very much like Ashby's Homeostat.

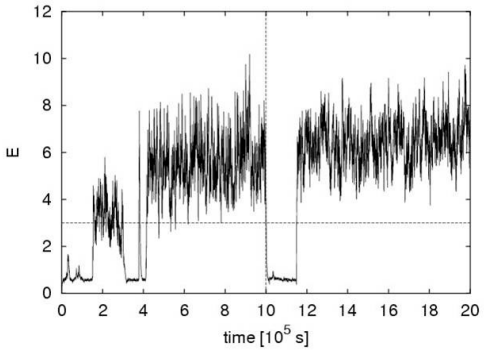


Figure 6. Battery level of robot: horizontal line indicates lower boundary (upper boundary was set to 20) and vertical line the time of visual inversion.

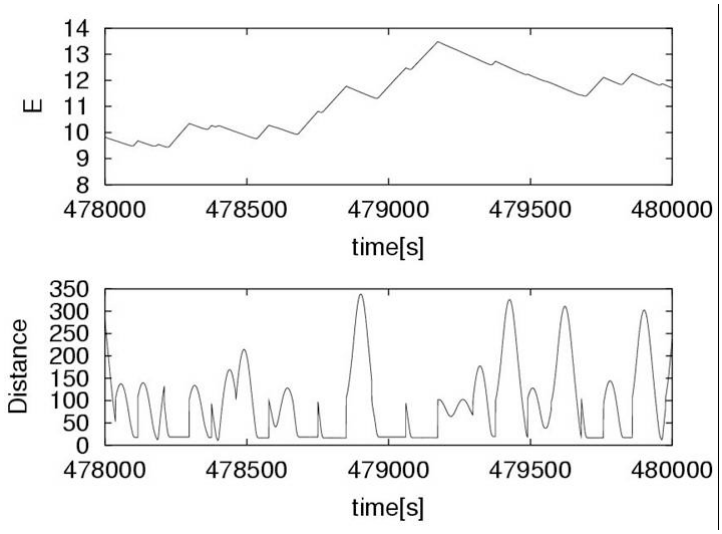


Figure 7. Battery level (top) and distance to source (bottom) in a phototactic robot.

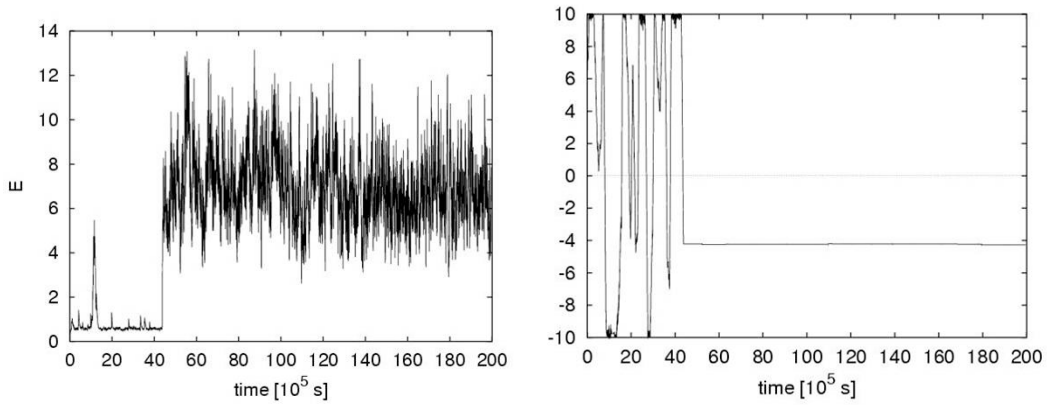
Although it will eventually adapt to almost anything, these experiments are helpful in identifying two other problems:

4. There are no guarantees regarding the time to adaptation.
5. There are no guarantees that adaptations will be conserved, in particular, that new adaptation will not interfere with previously acquired behaviours, if these are not re-evaluated with certain frequency.

In addition, the scheme makes such a direct connection between behaviour and internal dynamics that the result is almost (but not quite) trivial and the controller lacks interesting internal dynamics such as those that could be expected from a neural model.

Current explorations are looking at combining the first two schemes while trying to circumvent some of the problems that have been identified. For instance, to avoid the problem of loss of variability the plastic rules can be replaced by the following generic rule:

$$\frac{dw_{ij}}{dt} = \delta p_j \eta_{ij}(A_1(z_i - 0.5) + A_2(z_j - 0.5) + A_3(z_i - 0.5)(z_j - 0.5)),$$



(a) (b)

Figure 8. Sensors are swapped with the presentation of each new light source. The robot eventually adapts to both conditions at the same time: (a) plot of battery level, (b) illustration of parameter change (offset y_0 for connection between left sensor and right motor).

where the parameters A_k and η_{ij} are set genetically. Initial weight values are random and neuron activation is initialised randomly near the middle of the range (0.5) therefore also guaranteeing a random initial derivative.

Plastic facilitation p is changed to a symmetric function with continuous derivative which is only zero for a neural activation $z = 0.5$ and given by $p = (2z - 1)^2$. The damping factor δ now depends on the direction of change as well as the current weight value. When potentiating a synapse δ goes linearly from 1 at the minimum of the weight range to 0 at the maximum and inversely when depressing. This avoids the problems of weights "getting stuck" at the minimum and maximum values and leads to more uniform weight distributions.

In addition to these changes, the scheme can be made similar to the second experiment in that environmental conditions directly affect the internal stability of the neural controller so a dynamical link between the task and homeostasis, the necessary two-way causal relation, obtains. This may be achieved, for instance, by making the level of neural noise dependent on the distance to the source in an attempt to address issue 3 above.

It is clear that the issues 1 to 5 still need to be further investigated, and we do not pretend that these experiments in homeostatic adaptation are the only (or even a very good) way of exploring the points raised in the rest of the paper. But they do provide one possible way in, and some idea of what the required methodologies might look like, and so they are presented here only in this spirit.

7. Conclusions

In time, it was perhaps inevitable that artificial intelligence should turn to biology. Observing the richness of animal behaviour provides us with inspiration and aspirations for robotics, and it is only natural that we should start by imitating only selected properties of biological systems. It is definitely not the message of this paper that this practice should be abandoned. On the contrary, it should be pursued with more seriousness. We should aspire to imitate the *principles* of the biological world (as opposed to imitating

only what biologists and neuroscientists are currently working on). This is why asking the difficult questions is important.

In this paper we have done the following:

1. indicated what current robots are missing: true intentionality, true autonomy, and true teleology;
2. indicated why (and how), organismic inspiration, more that biological inspiration may be the answer to these problems;
3. identified intrinsic teleology as originating in the metabolising, autopoietic organisation of life;
4. explained how the viability condition has been put in terms that are very useful for robot design in the work of Ashby;
5. analysed why robots do not exhibit natural teleology or intentionality, in spite of having closed sensorimotor feedback loops;
6. specified how such loops must be complemented with a two-way link between internal organisation and behaviour for robots to better approach natural agents;
7. posed the question of whether intentional robots need to metabolise, or whether other principles might be applied.
8. proposed an answer to this question – robots need not metabolise, the principles of habit formation provide a better point of entry for finding methods to be incorporated in robot design; however, a better theoretical understanding is badly needed in this area; and
9. illustrated some preliminary ways of putting these ideas at work in the evolutionary design of homeostatically adaptive robots.

All of these issues deserve further development, as well as clearer methodologies for their practical application. They have been presented here in full knowledge of this fact. However, we are convinced of the value of raising the hard questions (and the occasional half-baked answer) for generating discussion. If this paper raises awareness of the current limitations in biologically-inspired robotics, and helps to open up the debate, its objectives will have been accomplished.

References

Ashby, W. R. (1947). The nervous system as a physical machine: with special reference to the origin of adaptive behaviour. *Mind*, **56**, 44–59.

Ashby, W. R. (1960). *Design for a Brain: The Origin of Adaptive Behaviour* (Second edition). London: Chapman and Hall.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, **4**, 91–99.

Beer, R. D., Chiel, H. J., Quinn, R. D., & Ritzmann, R. E. (1998). Biorobotic approaches to the study of motor systems. *Current Opinion in Neurobiology*, **8**, 777–782.

Bernstein, N. (1967). *The coordination and regulation of movements.* New York: Pergamon Press.

Braitenberg, V. (1984). *Vehicles: experiments in synthetic psychology.* Cambridge, MA: MIT Press.

Breazeal, C. (2001). Affective interaction between humans and robots. In Kelemen, J., & Sosík, P. (Eds.), *Advances in Artificial Life: Proceedings of the Sixth European Conference on Artificial Life*, pp. 582 – 591. Springer Verlag.

Brooks, R., Breazeal, C., Marjanovic, M., Scassellati, B., & Williamson, M. (1999). The Cog Project: Building a Humanoid Robot. In Nehaniv, C. (Ed.), *Computation for Metaphors, Analogy, and Agents, Lecture Notes in Artificial Intelligence 1562*, pp. 52–87. New York, Springer-Verlag.

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

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

Cliff, D. (1991). Computational Neuroethology: A Provisional Manifesto. In Meyer, J.-A., & Wilson, S. (Eds.), *From Animals to Animats, Proc. of first Intl. Conf. on Simulation of Adaptive Behavior*, pp. 29 – 39. Cambridge, MA: MIT Press.

Collins, S. H., Wisse, M., & Ruina, A. (2001). A 3-D passive-dynamic walking robot with two legs and knees. *International Journal of Robotics Research*, **20**, 607–615.

Dewey, J. (1896). The reflex-arc concept in psychology. *Psychol. Review*, **3**, 357 – 370.

Dewey, J. (1922). *Human Nature and Conduct. John Dewey, The Later Works 1925-1953, V.4.* Southern Illinois University Press, 1983.

Dewey, J. (1929). *Experience and Nature* (Second edition). New York: Dover. (1958).

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* Paris, France. Cambridge MA: MIT Press.

Di Paolo, E. A. (2003). Spike-timing dependent plasticity for evolved robot control: neural noise, synchronisation and robustness. *Adaptive Behavior*, **10(3/4)**. Forthcoming.

Di Paolo, E. A., Noble, J., & Bullock, S. (2000). Simulation models as opaque thought experiments. In Bedau, M. A., McCaskill, J. S., Packard, N. H., & Rasmussen, S. (Eds.), *Artificial Life VII: The Seventh International Conference on the Simulation and Synthesis of Living Systems*, pp. 497 – 506. Cambridge MA: MIT Press.

Dreyfus, H. L. (1979). *What computers can't do:The Limits of Artificial Intelligence* (Revised edition). New York: Harper and Row.

Floreano, D., & Mattiussi, C. (2001). Evolution of spiking neural controllers for autonomous vision-based robots. In Gomi, T. (Ed.), *Evolutionary Robotics IV*. Springer Verlag.

Floreano, D., & Urzelai, J. (2000). Evolutionary Robots with on-line self-organization and behavioral fitness. *Neural Networks*, **13**, 431 – 443.

Fontana, W., & Buss, L. (1996). The barrier of objects: from dynamical systems to bounded organizations. In Casti, J., & Karlqvist, A. (Eds.), *Boundaries and Barriers*, pp. 56 – 116. Addison-Wesley, Reading, MA.

Goldstein, K. (1995/1934). *The Organism.* New York: Zone Books.

Goodwin, B. C., & Webster, G. (1997). *Form and Transformation: Generative and Relational Principles in Biology.* Cambridge University Press.

Grene, M. (1968). *Approaches to a philosophical biology.* Basic Books.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, **42**, 335 – 346.

Harrington, A. (1996). *Reenchanted science: holism in German culture from Willhelm II to Hitler.* Princeton University Press.

Harvey, I., P., H., Cliff, D., Thompson, A., & Jakobi, N. (1997). Evolutionary robotics: the Sussex approach. *Robotics and Autonomous Systems*, **20**, 207 – 224.

Horn, D., Levy, N., & Ruppin, E. (1998). Memory maintenance via neuronal regulation. *Neural Computation*, **10**, 1–18.

Husbands, P., Smith, T., Jakobi, N., & O'Shea, M. (1998). Better living through chemistry: Evolving GasNets for robot control. *Connection Science*, **10**, 185–210.

James, W. (1890). *The Principles of Psychology*. New York: H. Holt and Company.

Jonas, H. (1966). *The phenomenon of life: towards a philosophical biology*. Northwestern University Press.

Kohler, I. (1964). The formation and transformation of the perceptual world. *Psychological Issues*, **3**, 1–173.

Lenoir, T. (1982). *The strategy of life: Teleology and mechanics in nineteenth century German biology*. The University of Chicago Press.

Maturana, H., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht, Holland: D. Reidel Publishing.

Maturana, H., & Varela, F. J. (1987). *The tree of knowledge: The biological roots of human understanding*. Boston, MA: Shambhala.

McGeer, T. (1990). Passive dynamic walking. *International Journal of Robotics Research*, **9**, 62–82.

McMullin, B., & Varela, F. J. (1997). Rediscovering Computational Autopoiesis. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life*, pp. 38 – 47. MIT Press, Cambridge, Mass.

Merleau-Ponty, M. (1942/1963). *The structure of behaviour*. London: Methuen. Translated by A. L. Fisher.

Millikan, R. G. (1984). *Language, thought and other biological categories: New foundations for realism*. Cambridge: MIT Press.

Nagel, E. (1977). Teleology revisited. *Journal of Philosophy*, **76**, 261–301.

Nolfi, S., & Floreano, D. (2000). *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*. Cambridge MA: MIT Press.

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

Piaget, J. (1948). *La naissance de l'intelligence chez l'enfant*. Delachaux et Niestlé, Neuchátel-Paris.

Piaget, J. (1967). *Biologie et Connaissance: Essai sur les relations entre les régulations organiques et les processus cognitifs*. Gallimard.

Rasmussen, S., Baas, N. A., Mayer, B., Nilsson, M., & Olesen, M. W. (2001). Ansatz for dynamic hierarchies. *Artificial Life*, **7**, 329 – 353.

Reeke Jr., G. N., Sporns, O., & Edelman, G. M. (1990). Synthetic neural modeling: the "Darwin" series of recognition automata. *Proc. of the IEEE*, **78**, 1498–1530.

Rosen, R. (1991). *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life*. Columbia University Press.

Rosenblueth, A. N., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy og Science*, **10**, 18–24.

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*, pp. 290 – 298. MIT Press, Cambridge, Mass.

Salomon, R. (1998). Achieving robust behavior by using proprioceptive activity patterns. *BioSystems*, **47**, 193–206.

Scheier, C., & Pfeifer, R. (1995). Classification as sensory-motor coordination: A case study on autonomous agents. In Moran, F., Moreno, A., Merelo, J. J., & Chacon, P. (Eds.), *Proceedings of the 3rd European Conference on Artificial Life*, pp. 657 – 667 Granada, Spain. Springer Verlag, Berlin.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, **3**, 417 – 457.

Stewart, J. (1996). Cognition = life: Implications for higher-level cognition. *Behavioural Processes*, **35**, 311 – 326.

Straus, E. (1966). *Phenomenological psychology*. London: Tavistock Publications.

Taylor, C. (1964). *The explanation of behaviour*. London: Routledge and Kegan Paul.

Taylor, J. G. (1962). *The Behavioral Basis of Perception*. New Haven: Yale University Press.

Turrigiano, G. G. (1999). Homeostatic plasticity in neuronal networks: The more things change, the more they stay the same. *Trends Neurosci.*, **22**, 221–227.

Uexküll, J. (1934). A stroll through the worlds of animals and men. In Lashley, K. (Ed.), *Instinctive behavior*. International Universities Press.

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

Varela, F. J. (1995). The re-enchantment of the concrete. In Steels, L., & Brooks, R. (Eds.), *The Artificial Life Route to Artificial Intelligence*. Hove, UK: Lawrence Erlbaum.

Varela, F. J. (1997). Patterns of life: intertwining identity and cognition. *Brain and Cognition*, **34**, 72 – 87.

Varela, F. J., Maturana, H. R., & 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.

Wehner, R., Michel, B., & Antonsen, P. (1996). Visual navigation in insects: coupling of egocentric and geocentric information. *Journal of Experimental Biology*, **199**, 129 – 140.

Wheeler, M. (1997). Cognition's coming home: the reunion of life and mind. In Husbands, P., & Harvey, I. (Eds.), *Proceedings of the Fourth European Conference on Artificial Life*, pp. 10 – 19. MIT Press, Cambridge, Mass.

Williamson, M. M. (1998). Neural control of rhythmic arm movement. *Neural Networks*, **11**, 1379–1394.

Wright, L. (1976). *Teleological Explanations*. University of California Press.