Ago Ergo Sum*

Dario Floreano
Evolutionary & Adaptive Systems Team
Institute of Robotics
Swiss Federal Institute of Technology at Lausanne (EPFL)
CH-1015 Ecublens, Switzerland
Dario.Floreano@epfl.ch
http://dmtwww.epfl.ch/isr/east/

Abstract

In this paper I explore the hypothesis that some of today's robots might possess a form of consciousness whose substrate is a mere algorithm. First, consciousness is defined within an evolutionary framework as awareness of one's own state in relation to the external environment. Then, the basic prerequisites for such conscious activity are discussed, namely embodiment, autonomy, and adaptation mechanisms. Artificial evolution, rather than evolutionary optimisation, is presented as a viable methodology to create conscious robots, accompanied by some examples of behaviours of artificially evolved robots. Finally, I argue that what might be problematic with the concept of robot consciousness is not the robot, but the notion of consciousness itself.

1 Paving the Road to Robot Consciousness

Could the ordered list of operations which forms any computer algorithm constitute the basis of consciousness for a robot? "Of course not", would be the sensible answer probably given by most readers with a scientific or engineering oriented mind. The popular justification for such a view is that computer programs are instructions used to control machines which must carry out a specific task, such as a lengthy computation or assembly of car components. A robot controlled by such an algorithm would be nothing more than an automaton, a mechanical device which blindly and repetitively executes a pre-specified set of instructions (see [39, chapter 1] for elegant elaborations of this view). In this paper, I shall consider a different position and argue that it is not possible to exclude that some of today's robots might indeed possess a form of *proto-consciousness* whose substrate is a mere algorithm, or something that can in principle be described in terms of algorithmic computation (e.g., an artificial neural network, a rule-based system, a classifier system, etc.). I will do so by

[*] *Ago*, first person singular of the Latin verb of Greek root, meaning "to lead, to play an active role, to behave". This article was originally written in 1996 under invitation, but recent research advancements do not contradict its contents.

showing why this might be possible at all and by describing the basic ingredients that, if accurately mixed together, might give rise to forms of conscious behaviour. I will support my line of reasoning with examples of behaviours recorded during experiments of artificial evolution performed on mobile robots at our laboratory.

All of what follows assumes that consciousness exists, if not as a reality, at least as a scientific hypothesis worth considering. Before proceeding any further, though, it is necessary to explain exactly what I mean by consciousness here in order to make my ideas testable and concretely debatable.

1.1 Consciousness as Awareness

Consciousness is difficult to define in scientific terms because it has many facets and attempts to provide an exaustive description often involve a good deal of introspection. Whatever consciousness might be, it certainly is a product of natural evolution, the result of several generations of selective reproduction. Given the fierce competition for energy and resources that characterises all aspects of life, it is reasonable to assume that consciousness provided some survival advantage for organisms that began to exhibit it, even though its primordial manifestations migh have been an interesting byproduct of evolution with no negative impact on inclusive fitness. If we consider consciousness as an ability for better coping with the difficulties faced by a living organism, then it becomes inevitable that its characterisation should bear some relationship to the basic challenges on which survival and success of reproduction is based. These include, in increasing order of complexity (from [8]): sensori-motor coordination and proprioception, object recognition and recognition of spatio-temporally invariants, spatial navigation, multi-modal sensory fusion and logical processing, social behaviour. The common factor underlying all these abilities is the required capacity of an organism to position itself with respect to the external environment, that is to achieve increasingly complex levels of situatedness.

Therefore, among all the possible definitions, here consciousness is taken to be *awareness* of one's own internal state with respect to the environment. Being conscious is the process by which an intelligent control system performs spontaneous self-monitoring of internal states (which could take into account not only neurally generated activity, but also physiological states) by putting them in relation to the external environment. Thus, although consciousness is an internal reflexive activity, it is not purely subjective and disjointed from the external environment (which would rather correspond to hallucination), but is intimately linked to purposeful behaviour. Within this framework, consciousness becomes useful because it allows comparison and processing of information coming from several sensory modalities and it provides the system with the ability to make predictions in order to change its course of actions. By spontaneously monitoring, cross-correlating, and variously arranging several internal processing states in relation to what happens in the environment, one can anticipate different behavioural outcomes and act accordingly (see also [10, chapter 7]). This type of consciousness also requires active exploration. When the system realises that there is not sufficient information for performing an action, or feedback from the environment does not correspond to the internally anticipated situation, the organism will engage in exploratory behaviour in order to actively seek missing information or in order to change some of its own internal parameters. I think that this type of consciousness is one of the characteristics that distinguishes purely reactive organisms from intelligent organisms. In simple words, consciousness is a way of continuously refreshing and updating one's position in the world, a way of making sense of what is happening by coordinating relevant information and of anticipating future situations for appropriate action.

The definition given above is quite reductive, but I think that it captures the essential elements of what consciousness is, the basic purposes of conscious activity in humans as well as possibly in other organisms. To this extent, my analysis of consciousness as awareness of one's own state refers to a type of "proto-consciousness", similar for several aspects to the definition of "primary consciousness" given by Edelman [13], which does not exclude other phenomena, such as feelings, sensations, etc., as long as they participate to the goal of establishing and maintaining the position of the organism in its own environment. However, I deliberately do not consider phenomenal consciousness here, which is largely based on introspection and might be characteristic of human beings only or a pure artifact of our reasoning abilities.

In the remainder of this paper, I will try to show that a careful consideration of some evolutionary issues might indeed give us a methodology to create with our current technology conscious—that is, aware—artificially living organisms.

2 Prerequisites

One reason why attributing consciousness to an algorithm sounds weird is that we tend to see algorithms as abstract instruction sets designed for a specific computational purpose. I agree that such an algorithm cannot exhibit any form of awareness. The type of interactions between such an algorithm and the external world —if any— is precisely defined and regulated by the conditional statements (e.g., `if-else`) which form the algorithm itself and guide its operation toward the achievement of a pre-specified goal. There are no reasons whatsoever why such an algorithm should be aware of its position in the environment, change its course of action or its processing modality.

2.1 Embodiment

Things might change when the algorithm becomes *embodied*, that is when it is part of a body equipped with a sensori-motor system; now the algorithm is called a "control system". A body is a physical entity which interacts with a physical environment in forms which might drastically affect the processing modality of its own controller. For example, let us consider information processing in a certain type of algorithm, an artificial neural network which maps input vectors into appropriate output vectors. Here the sequence of input-output vectors (both during the learning phase and during normal functioning) is quite arbitrary, usually random or in accordance with a certain schedule which has been carefully chosen by the external user in order to maximise a well-defined objective function. Instead, when the neural network is embodied in a sensorimotor system (wherin it is often called a neurocontroller), the *time sequence* of the input vectors (sensory information) depends on the output (motor actions) of the network itself, which in turn depends on the input. In this case the flow of information between the network and the environment is coherent, meaningful, and self-contained (in the sense that it is partially or completely generated by the information-processing system itself) [38]. The interaction with the environment affects not only the time sequence, but also the *type* and *frequency* of information which is available to the network. That means that the type of experiences to which such a network is exposed greatly depends on the network itself. Such a neural network could actively avoid or seek certain situations as well as potentially affect its own learning modality. If the control system was not part of a physical body struggling for survival, all this potential freedom could degenerate in un-interesting behavioural loops where the system closes on itself and minimises interactions with the external world. However, bodies have physiological demands—such as keeping an adequate temperature level and maintaining a sufficient amount of energy—which continuously urge the organism to develop better and more efficient behavioural strategies in order to satisfy them.

If we sum up the consequences of embodiment mentioned above, it becomes evident that such a control system would gain considerable advantage from the type of conscious activity described in section 1.1, that is the ability to monitor and regulate internal and physiological states; to seek, integrate, and coordinate the multi-modal flow of information in order to establish appropriate correlations, highlight important bits, and suppress the rest; to distinguish between environmental changes that are due to its own actions and those that are due to external causes (in other words, to understand the sensory consequences of its own actions [38]); to anticipate future situations in order to find a timely solution and avoid dangerous situations. These considerations apply both to living organisms and to artificial organisms, such as robots where the physiological demands might correspond to the ability to maintain an appropriate reserve of electrical power or to minimise mechanical wear.

2.2 Autonomy

However, embodiment alone is not sufficient to make us believe that some type of computation, be it biological or artificial, might exhibit consciousness. After all, there are several examples of robots that are precisely pre-programmed to respond in precise ways to incoming information; e.g., the already mentioned robots working on industrial assembly lines. In that case, a human analyst designs the control system of the robot so that it performs exactly and precisely certain actions in response to a restricted class of environmental conditions in order to carry out a pre-defined task. Here, the robot does not have freedom of exploration, it cannot establish its own goals and sub-goals, or change the way in which it acts; in other words, such a robot would not display any awareness because the flow of information from sensors to actuators has been rigidly specified in detail by an external designer for the purpose of achieving an externally defined goal. In fact, another important pre-requisite of conscious activity—both in living and artificial organisms—is *autonomy.* Autonomy implies freedom from external control [31]. An autonomous agent is a system that is capable of defining its own goals and regulating its own behaviour accordingly; in other words, it must be capable of self-control [44]. Autonomy is often associated with self-sufficiency, but the two notions are in fact quite different. For example, a bird raised in a cage is not self-sufficient, and if put in the wild it will probably die because it has not developed efficient foraging strategies; nevertheless, it is an autonomous agents. Similarly, a robot might be autonomous even though it depends on external provision of electrical power. The definition of autonomy is concerned with the locus of control, i.e. with who decides the goals of the agent, not with the capacity for self-sufficiency.

2.3 Adaptation

The notion of autonomy clashes with the traditional method of engineering behaviours for robots, which could be summarised in the following logical and temporal steps: **I.** Create a formal model of the robot and of the environment; **II.** Analyse the task requirements and decompose it in appropriate subtasks; **III.** Design an algorithm which steers the robot in accordance with the formal understanding gained in the previous steps. Such an agent could hardly be described as autonomous. As McFarland has put it [32, introduction], creation of autonomous agents requires a shift in behaviour engineering from a goal-directed to a goal-seeking and self-steering control system (see also [40]). An autonomous agent cannot be programmed in the traditional way described above. Rather, its control system should be equipped with *learning* mechanisms that give the agent the ability to adapt to the environment and to develop its own strategies in order to maintain itself in a viable state. A very popular choice consists in using neurocontrollers—artificial neural networks where the input units are clamped to the robot sensors and the output units control the effectors, such as wheels, grippers, etc.—which are well suited structures for learning mechanisms (for examples, see some recent special journal issues on robot learning [20, 11]). Learning consists in a gradual modification of the internal parameters (synaptic connection strengths, neuron thresholds, and architecture) while the robot interacts with the environment. However, one might envision a variety of different processing metaphors (such as classifier systems) that are capable of autonomous self-configuration. What really matters in these adaptation techniques is to what extent learning proceeds according to an externally defined goal and schedule rather than resulting from a process of self-organisation. These two learning modalities are often referred to as supervised and unsupervised, respectively (e.g., see [24]). Probably, it is impossible to draw a sharp line between them because in both cases adaptation takes place in order to optimise an objective function (e.g., a cost function or an energy function), which seems to hold for learning and behaviour regulation in living systems too [32]. The difference between learning automata and learning autonomous agents, then, lies in the type of information which is made available to the agent for learning and in the way it is used. For example, modifying the synaptic strength by backpropagation of error [43] between the agent's actions and the actions defined by a human observer, is an efficient—when feasible—way of programming automata (e.g., see [41]); there, learning is aimed at achieving a specific goal by implementing a detailed behavioural schedule predefined by an external user. Instead, a more suitable technique for creating autonomous agents might be provided by reinforcement learning (for a clear review, see [3]), where learning takes place only when the agent receives a reward signal from the environment. Reinforcement learning algorithms try to maximise the probability of repeating behaviours which lead to positive reinforcement signals. Despite being a promising approach, these algorithms require an extensive initial exploration of the possible behavioural outcomes in order to develop stable controllers, which does not seem always to be the case in living organisms.

In this paper, I will consider a different adaptation mechanism, *artificial evolution of neurocontrollers*, as a viable methodology to develop autonomous agents that could exhibit conscious abilities. Artificial evolution differs from other learning schemes because it works on a population of different individuals and is based on a selectionist, rather than goal-directed, approach [44].

3 Artificial Evolution of Autonomous Robots

Application of evolutionary techniques for automating the development of robotic control systems has been tried with success by several researchers (see [30] for an extensive review). Also referred to as "Evolutionary Robotics" [9], this approach relies on at least two interesting issues: **a)** the possibility of replicating and understanding basic principles of adaptation in artificial forms of life [28], and **b)** the generality and the power of evolutionary techniques to develop control systems exhibiting complex behaviours which would otherwise be difficult to design by hand. In several cases, the basic building blocks specified in the artificial genomes are networks of real-time recurrent neurons, which are parallel processing systems well-suited for operating in real-world conditions [23] and for potentially exhibiting complex and minimally cognitive behaviours [45, 4]. Although the specific algorithmic instantiation of the control system is not crucial for creating autonomous agents, the definition of consciousness given above requires some type of feedback and lateral flow of information among the parallel processes that compose the control system. These information channels, which are easily implemented as recurrent connections in artificial neural networks, might serve the purpose (often rethorically attributed to an internal *homunculus*) of establishing appropriate multi-modal correlations, of coordinating internal states, of monitoring and changing the course of actions on the basis of previous experience, and of anticipating future internal and sensory states (see also [14] for an extensive review of the relevance of "re-entrant connections" in living and artificial conscious systems).

Recently, the well-known issue of agent simulation versus physical implementation has been re-examined with respect to evolutionary robotics. All evolutionary techniques operate and rely on populations of different individuals, several of which display maladaptive behaviours, especially in the initial generations. This represents a heavy burden for physical implementations on real robots in terms of time and resources. Therefore, most of the experiments of evolutionary robotics are carried out in simulations, and only in few cases is the evolved controller transferred on the real robot (e.g., see [37, 33]). This procedure has been criticised because it might not scale well as soon as the robot geometry and dynamics become more complex [30]. Although for the purpose of the arguments offered in this paper it does not really matter whether the agent is simulated or physically implemented, all the experiments that I will report below have been entirely carried out on a real mobile robot.

3.1 Artificial Evolution vs. Evolutionary Optimisation

Computationally speaking, evolutionary methods represent a very general technique because they can be applied to any aspect of the agent (synaptic connections and number of neurons [46], growing factors [6], body specifications [29], etc.) and to any type of algorithm instantiation (neural networks [42], classifier systems [12], graph systems [26], programming language functions [27], etc.) as long as they can be mapped into a genetic description. It has been empirically shown that genetic algorithms outperform other search methods when the space of the possible solutions is high-dimensional and non-differentiable (see [21, chapter 1] for a simple description of where and why evolutionary search outperforms other search methods). These two properties, generality and efficiency, are often exploited to optimise unknown parameters of a complex system so that it will exhibit a well-defined behaviour (e.g., see [2]). To the extent that behaviour generation is externally defined and driven to a specific goal, evolutionary optimisation is not very different from the supervised learning approach described in section 2.3. However, natural evolution does not have the notion of teleology which is so familiar to biologists [1] and engineers, but is rather an open-ended process (see also [22]) where the concept of optimum cannot be universally defined. It would be wrong to assume that a specific organ, such as the retina, has evolved in order to maximise a specific function, such as transmission of reflected light. Similarly, there are no goals driving evolutionary development of cognitive systems; where goals exist, they are autonomously created by the agent itself as a more efficient way of organising its own survival strategies. If taken to its extreme consequences, this reasoning would imply that artificial evolution should not make use of any externally defined fitness function and perhaps it should substitute selective reproduction with elimination of individuals which cannot keep themselves viable. However, also in this case, it would be inevitable to introduce some external bias by deciding what viability is in the specific artificial implementation of the ecosystem. Indeed, it is very difficult to draw a line between evolutionary optimisation and artificial evolution. With respect to the issue of evolvability of conscious control systems, we can re-state these differences by saying that the more consistent is the shift from evolutionary optimisation to artificial evolution, the bigger is the probability that the evolved controllers exhibit consciousness.

In the next sections, I will try to support this statement by describing two experiments of evolutionary robotics where more complex neurocontrollers can be evolved by simply lifting some constraints on the fitness function and considering the robot as an organism with its own physiological requirements. Although still simple, these controllers exhibit several of the computational and behavioural characteristics listed above as indicators and prerequisites of conscious activity.

3.2 Some Experiments in Evolutionary Robotics

The experiments described here have been carried out on a real mobile robot without human intervention. In all cases we employed a single physical robot which served as body for several populations of individuals which were tested one by one. Although this procedure is biologically implausible because in biological life control systems and bodies co-evolve, hardware evolution is still a challenging technical issue (see [29] for some initial attempts in this direction). Each individual control system was a neural network where the input units were clamped to the robot sensors and the output unit activations were used to set the rotation speed of the wheels. Different neural architectures were used in the various experiments, but they all had sigmoid activation function and recurrent

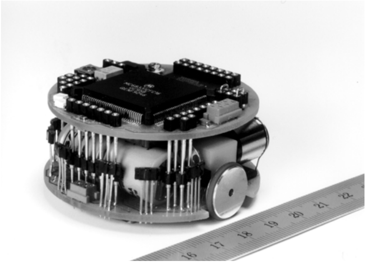

Figure 1: Khepera, the miniature mobile robot used for the experiments of artificial evolution.

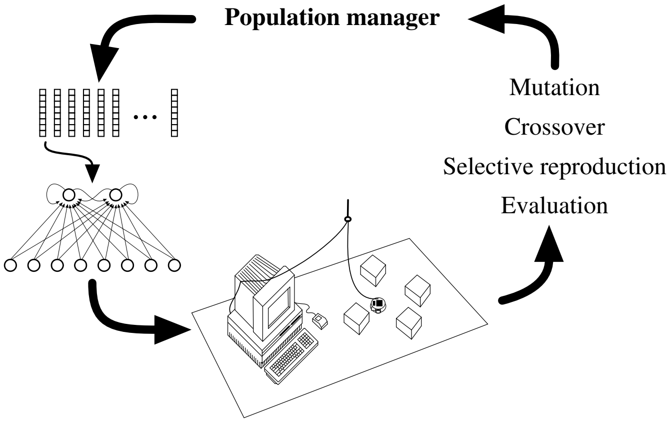

Figure 2: Operating methodology.

connections.

The robot employed in the experiments is Khepera, a miniature mobile robot [35]. It has a circular shape (Figure 1), with a diameter of 55 mm, height of 30 mm, and weight of 70 g, and is supported by two wheels and two small Teflon balls. The wheels are controlled by two DC motors with incremental encoder (10 pulses per mm of advancement of the robot), and can rotate in both directions. In the basic configuration, the robot is provided with eight infrared sensors, six positioned on one side of the robot (front), the remaining two on the other side (back). Each sensor can function in two modalities: it can emit and measure reflected infrared light and it can measure the infrared light component of the ambient light. When functioning in the first modality, it is called a proximity sensor because it is more active when it is closer to an object; when functioning in the second modality, it is called a light intensity sensor because it gives a rough indication of the ambient light intensity at its location. A Motorola 68331 controller with 256 Kbytes of RAM and 512 Kbytes ROM manages all the input-output routines and can communicate via a serial port with a host computer.

Khepera was attached via a serial port to a Sun SparcStation 2 by means of a lightweight aerial cable and specially designed rotating contacts (see [34] for more detailed descriptions of technical issues related to the experiments described in this paper). In this way we could exploit the power and disk size available in a workstation by letting high-level control processes (genetic operators, neural network activation, variables recordings) run on the main station while low-level processes (sensor-reading, motor control, and other real time tasks) run on the on-board processor (Figure 2). Thus, while the robot was operating, we could keep track of all the populations of organisms that were born, tested, and passed to the genetic operators, together with their "personal life files". At the same time, we could also take advantage of specific software designed for graphic visualization of trajectories and sensory-motor status while the robot was evolving [7]. As already stated in section 2.2 the fact that the robot depended on an external source of energy does not affect its autonomy. Also, for what concerns Khepera, the robot is not aware of where its own "brain" is located as long as it is connected to its own sensors and motors.[1]

The evolutionary procedure was a standard genetic algorithm as described by Goldberg [21] with fitness scaling and roulette wheel selection, biased mutations [36], and one-point crossover. Each individual of a population was in turn decoded into the corresponding neural networks, the input nodes connected to the robot sensors, the output nodes to the motors, and the robot was left free to move for a given number of steps (motor actions) while its fitness Φ was automatically recorded. Each motor action lasted 300 ms. Between one individual and the next, a pair of random velocities was applied to the wheels for 5 seconds.

3.3 Automatic Evolution of Behavior

The first experiment was chiefly aimed at testing the evolutionary approach on a real mobile robot and assessing advantages and difficulties (see [17] for more details). We decided to evolve a neural network with a single layer of weights to perform straight navigation and obstacle avoidance. The robot was put in an environment consisting of a sort of circular corridor whose external size was approx. 80x50 cm large (Figure 3). The walls were made of light-blue polystyrene and the floor was a thick gray paper. The robot could sense the walls with the IR proximity sensors. Since the corridors were rather narrow (8-12 cm), some sensors were slightly active most of the time. The environment was within a portable box positioned in a room always illuminated from above by a 60-watt bulb light. A serial cable connected the robot to the workstation in our office a few rooms away.

The fitness function Φ was explicitly designed for straight navigation and obstacle avoidance

$$\Phi = V\left(1-\sqrt{\Delta v}\right)(1-i) \tag{1}$$

$$\begin{array}{rcccl} 0 & \leq & V & \leq & 1 \\ 0 & \leq & \Delta v & \leq & 1 \\ 0 & \leq & i & \leq & 1 \end{array}$$

[1] The software implementing the genetic operators and the neurocontroller [15] could be easily slimmed down and downloaded into the robot processor.

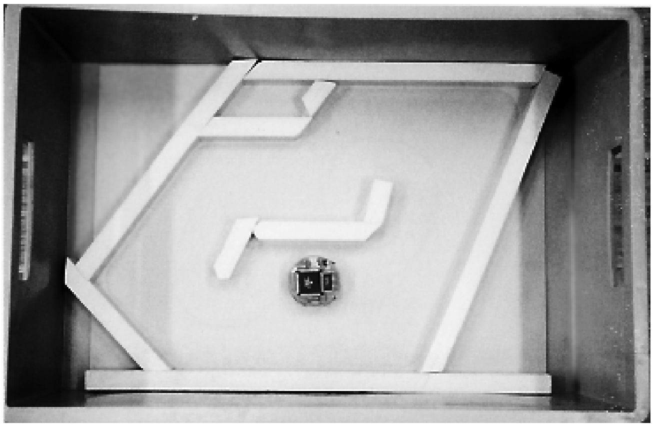

Figure 3: Environment of the experiment.

where V is a measure of the average rotation speed of the two wheels, Δv is the absolute value of the algebraic difference between the signed speed values of the wheels (positive is one direction, negative the other), and i is the activation value of the proximity sensor with the highest activity. The function Φ has three components: the first one is maximized by wheel speed (but not direction of wheel rotation), the second by straight direction, and the third by obstacle avoidance. The second component was very important. Without it, a robot could maximise its own fitness by simply spinning on itself at very high speed on a location far from obstacles, a trivial solution which can be found by simply setting a very high threshold value for one of the motor neurons and a strong synaptic value from any of the sensors to the other motor neuron. The square root was introduced after a series of trials and errors and was aimed at emphasising penalisation of small differences between the two wheel rotations for that specific training environment. Without it, the controller would set in sub-optimal solutions consisting in large circular trajectories.

In less than 50 generations, corresponding to approximately 24 hours of uninterrupted evolutionary adaptation, the best neurocontrollers of the population displayed smooth trajectories around the arena without hitting the walls (Figure 4). Thanks to the generalisation properties of artificial neural networks, these neurocontrollers could navigate in any type of environment (in different light conditions and with obstacles of different reflectance). These results were replicated several times by restarting the experiment from new randomly initialised populations. The resulting controllers cannot be properly called autonomous, but evolution did autonomously find a number of interesting solutions to the problem of navigation. For example, although the robot had a perfectly circular shape and the wheels could rotate in both directions, all the best neurocontrollers developed a frontal direction of motion corresponding to the side with more sensors, which gave a better resolution of the sensory information permitting better obstacle detection.

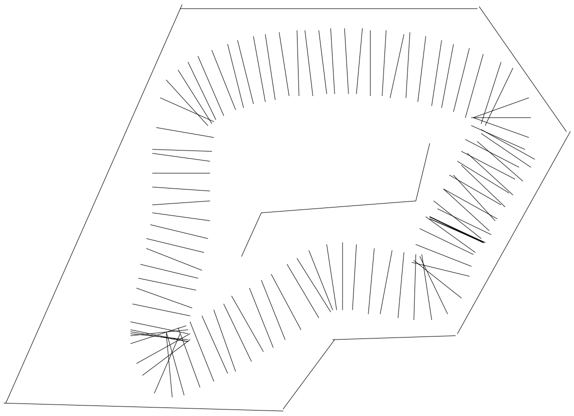

Figure 4: The trajectory performed by one of the evolved robots. Segments represent successive displacements of the axis connecting the two wheels. The direction of motion is anti-clockwise.

3.4 Evolution of an Autonomous System

Having ascertained that evolutionary techniques can be successfully applied to a real mobile robot, we decided to give the system more autonomy of development by applying the following changes:

- simplify the fitness function by eliminating the middle component;
- provide the robot with its own physiology, that is a limited life duration controlled by a rechargeable battery;
- make the environment more “interesting” by introducing a battery charger and a landmark.

The environment employed for evolutionary training consisted of a 40x45 cm arena delimited by walls of light-blue polystyrene and the floor was made of thick gray paper (Figure 5) as in the previous experiment. A 25 cm high tower equipped with 15 small DC lamps oriented toward the arena was placed in one corner. The room did not have other light sources. Under the light tower, a circular portion of the floor at the corner was painted black. The painted sector, which represented the recharging area, had a radius of approximately 8 cm and was intended to simulate the platform of a prototype battery charger. When the robot happened to be over the black area, its simulated battery became instantaneously recharged. The simulated battery was characterised by a fast linear discharge rate (max duration: approx. 20 seconds). In order to truncate the lives of individuals who would just sit on the battery charger or manage to recharge themselves, we set an upper life limit of 60 seconds for each robot, after which the next individual in the population was tested. By simulating the battery instead of using the onboard available batteries (which lasted much longer and also required much longer recharging time), artificial evolution produced

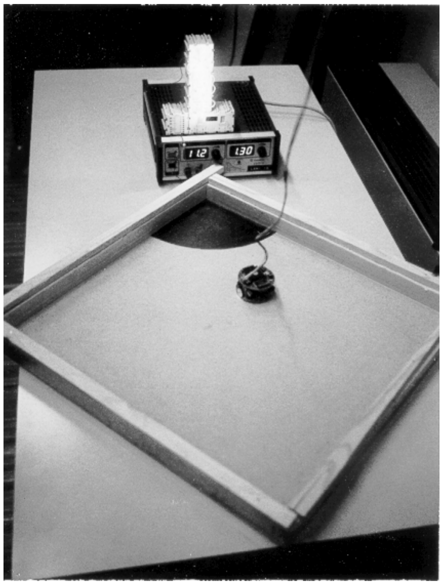

Figure 5: The environment of the experiment on battery recharge. The light tower is positioned in the far corner over the recharging area which is painted black. There are no other light sources in the room.

interesting results in approximately 240 hours instead of 6 years! The neurocontroller received input from the eight proximity sensors, from two ambient light sensors (each positioned on one side of the robot), from a detector of floor brightness placed underneath the robot platform, and from a sensor of battery level; these inputs projected to five internal units interconnected by lateral and recurrent connections, which then projected to two motor neurons controlling the wheels. The simplicity of the fitness function employed in this experiment may be exploited by a robot quickly spinning on the same spot far from the walls. The fitness function returned almost null values when the robot was on the battery charger (because it was positioned on a corner near the walls).

After ten days of continuous evolution (approx. 240 generations), the best neurocontroller of the population displayed interesting homing behaviour. By providing Khepera with a small "helmet" which gives us precise indication of where it is in the environment, we can correlate internal neural activity with its behaviour while it is autonomously moving in the environment. Figure 6 plots a typical trajectory of the robot (bottom-right) along with the activation of the five internal hidden units. Starting with a fully charged battery, the robot moves around the arena covering the whole area without hitting the walls and accurately avoiding the recharging zone. However, as soon as the battery level reaches a certain minimum value, the robot orients itself toward the recharging area and proceeds towards it on a straight line. Once recharged, it quickly exits the recharging area and resumes its large trajectories across the arena. Arrival to the recharging area is always extremely precise, approximately 1 second before total discharge. Since the trajectories around the environment are never the same, the robot autonomously decides what is the residual battery level when

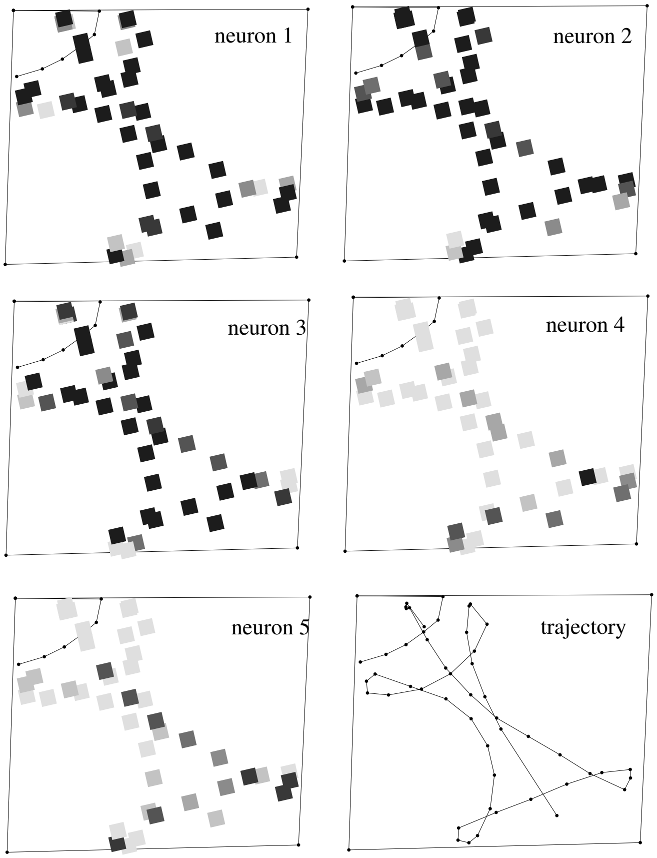

Figure 6: Visualisation of the activity of hidden units while the robot moves in the environment. Darker squares mean higher node activation. The robot starts in the lower portion of the arena. The bottom-right window shows the trajectory only. The recharging area is in the top left corner (Adapted from [18]).

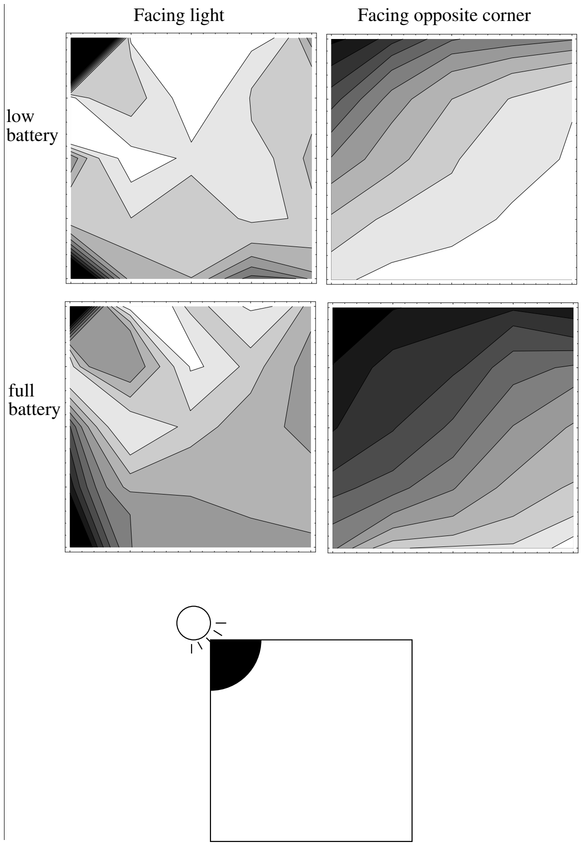

Figure 7: Contour map of the activation of neuron 5 for 2 different battery levels. The measurements were taken by positioning the robot at several evenly-spaced locations in the environment (depicted at the bottom). When the robot is facing the recharging area, the activity map does reflect the paths taken by the robot (Adapted from [18]).

it must move toward the recharging area, depending on where the robot is in the environment. When it is far away from the recharging area, Khepera will begin to turn and move toward it at a much higher residual level than when it is closer to the area. From these and other tests reported in [18], it becomes clear that the robot accurately avoids the recharging area if the battery level has not yet reached a certain minimum level, but it goes toward it if that level has been surpassed. By looking at the internal dynamics of the neurocontroller while the robot freely moves in the environment (Figure 6), we notice that neuron 5 becomes active shortly before the robot starts orienting toward the recharging area. Other neurons display significative changes of activity levels only when the robot is close to the walls and are thus responsible for the quite automatic behaviour of obstacle avoidance. By positioning the robot at various locations in the environment and measuring the activity of neuron 5 for a single instant, we can see that it has developed a map of the environment which varies depending on the orientation of the robot, but *not* on the battery level (Figure 7). Thus,

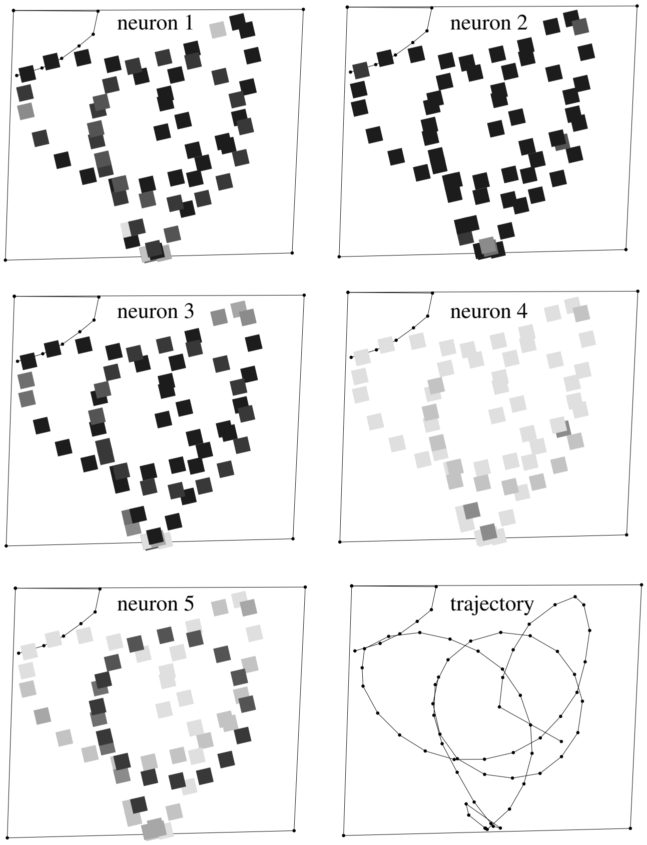

Figure 8: Visualisation of the activity of hidden units while the robot moves in the environment. Darker squares mean higher neuron activation. The robot starts in the lower portion of the arena. The bottom-right window shows the trajectory only. The recharging area is in the top left corner (Adapted from [18]).

neuron 5 creates an internal representation of where the robot is and correlates it with battery level information in order to switch the robot behaviour from simple navigation and obstacle avoidance into a precise and timely homing trajectory.

The switch from a rather automatic behaviour into an active “home-seeking” behaviour can be seen in a simple test where the robot is put in the environment with the light off. Now, its infrared proximity sensors can still detect the walls, but there is no more information for homing. Figure 8 shows the trajectory and neuron activations for this test. Initially, the robot is not disturbed by the dark environment and it performs the usual navigation and obstacle avoidance across the arena. As soon as the battery reaches a certain residual charge level (which now is much higher than in normal light condition), neuron 5 becomes active and the robot starts to perform a perfectly circular trajectory in the middle of the environment until it eventually stops operating. This circular path is a very smart exploratory strategy because, if there was some some weak light source in the arena, the robot could detect it and follow its gradient.

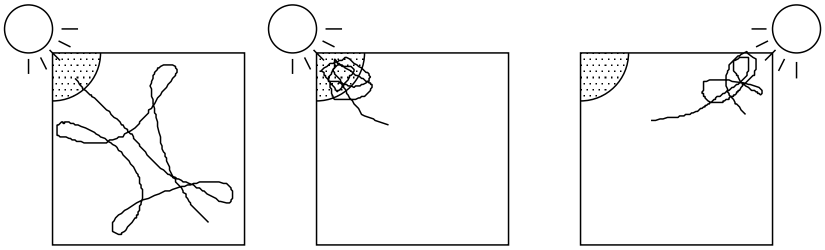

Figure 9: Trajectories of best individual of generation 240 in three environmental conditions. Left: Test in training conditions, from Figure 6. The robot starts with a full battery in the bottom right corner (only the first 50 actions are displayed). Centre: The robot starts in the centre of the environemnt with an almost discharged battery; the battery is not automatically recharged when the robot arrives on the charging area. Right: The light source is positioned on the top right corner, but the charging area remains at the original location.

We can do some further tests to check the reactions of the robot in unexpected situations (Figure 9). If the battery is not automatically recharged once the robot has arrived to the black area, it will continue to stay on it waiting for the recharge until it will eventually stop operating (Figure 9, centre). Similarly, if the light tower is moved to a different corner of the environment, but the recharging area remains in the same position, the robot will move toward the light; as soon as it does not detect the black painted surface, it will start searching in the surroundings until it will eventually stop operating (Figure 9, right).

Although re-adaptation to changes of landmark position can be easily achieved in a few generations of continued evolution [16], the system would greatly benefit from a combination of phylogenetic evolution and ontogenetic learning. The basic idea is that, instead of evolving neurocontrollers with fixed synaptic weights, it might be possible to evolve plastic networks that change their synaptic strengths according to evolved plasticity rules while the agent interacts with the environment. Preliminary successful results have already been achieved using the environment and fitness function described in section 3.3 [19], but they are not reported here because they were mainly aimed at assessing technical feasibility rather than at developing a fully autonomous agent.

None of the abilities and behaviours described above were explicitly described in the fitness function. However, the fact that longer life duration could correspond to the accumulation of more fitness points, evolutionary pressure selected individuals that:

- autonomously discovered the presence of the recharging area;
- correlated internal physiological requirements with the robot location in the environment to decide when to switch behaviour so as to avoid battery discharge;
- created an internal representation of the environment to perform efficient homing navigation;

- could substantially change their behaviour depending on their physiological state (battery level);
- autonomously decomposed the global behaviour in subgoals, i.e. automatic navigation with obstacle avoidance and homing navigation, which were managed by different internal processes;
- could actively search for missing information;
- displayed meaningful behaviours when “environmental reward” (battery recharge) was not available, in one case waiting for the expected recharge, and in the other case searching for the sensory cues associated with recharge.

These abilities reflect the type of activity and behavioural outcomes that were listed as indicators of conscious awareness in section 1.1; also, the algorithm that controls the robot was created in accordance with the prerequisites discussed in section 2.

4 What is Wrong with Artificial Consciousness?

In this paper, I have emphasised a specific aspect of consciousness, i.e. awareness of one’s own position with respect to the environment, using evolutionary and behavioural arguments; I have then argued that some of today’s robots might indeed exhibit such conscious ability even when their internal activity can be described by an algorithm, provided that the prerequisites of embodiment, autonomy, and adaptation are met. Finally, I have given some examples of artificially evolved neurocontrollers whose internal dynamics and corresponding behaviours match the requirements for conscious awareness. In my arguments, purposeful interaction with an environment plays a fundamental role. Whereas Penrose, who thinks that consciousness cannot be understood and replicated by means of current algorithmic computation, considers the environment as a potential source (quickly dismissed) of non-computability for an algorithmic system which would then raise the possibility of conscious phenomena [39, pp. 152–154], the environment here is important because it gives meaning to an otherwise abstract processing system and it drastically affects its processing modalities. Behaviour, not thinking, is the basis of conscious activity.

My conclusions are intentionally provocative. The essence of my argument is that if there is something wrong with the notion of consciousness in an algorithm-driven machine, the problem is not to be found in the algorithm or in the machine, but in the definition of consciousness. In all the definitions of consciousness I come across, there is always some element of introspection and subjectivism, if not poetry, which makes difficult any scientific conclusion on this issue and complicates the debate. Although here I have tried to take a definition of consciousness (or rather “proto-consciousness”) which could form the basis for debate or at least scientifically-motivated criticism, it is not clear whether one should introduce a new label like “consciousness” instead of simply using well-understood concepts like attention, cross-modal fusion, coordination, planning, etc.

What I have attempted to show is that current scientific and technological methods are sufficient to recreate artificial forms of life that display characteristics of biological intelligence. Therefore, it cannot be ruled out that these same artificial organisms could display forms of consciousness, whatever consciousness might be. After all, my conclusion is not a big conceptual advancement with respect to what Thomas Huxley said in 1874 during an invited address at the meeting of the British Society in Belfast under the title "On the hypothesis that animals are automata, and its history"

> One does not battle with drummers; but I venture to offer a few remarks for the calm consideration of thoughtful persons. It is quite true that, to the best of my judgement, the argumentation which applies to brutes [animals] holds equally good of men; and, therefore, that all states of consciousness in us, as in them, are immediately caused by molecular changes of the brain substance. [...] We are conscious automata, endowed with free will in the only intelligible sense of that much abused term—inasmuch as in many respects we are able to do as we like—but nonetheless parts of the great series of causes and effects which, in unbroken continuity composes that which is, and has been, and shall be—the sum of existence [from [25], partially reprinted in [5, p. 20]].

Acknowledgments

I would like to thank Francesco Mondada, Edo Franzi and André Guignard for designing and building the Khepera robot; I am especially grateful to Francesco Mondada for technical assistance and fruitful discussions on biologically inspired methods for engineering artificial systems.

References

[1] W. Atmar. Notes on the Simulation of Evolution. *IEEE Transactions on Neural Networks*, 5:130–148, 1993.

[2] S. Baluja. Evolution of an Artificial Neural Network Based Autonomous Land Vehicle Controller. *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, 26:450–463, 1996.

[3] A. G. Barto, R. S. Sutton, and C. J. C. H. Watkins. Learning and sequential decision making. In M. Gabriel and J. W. Moore, editors, *Learning and Computational Neuroscience*, pages 539–602. MIT Press-Bradford Books, Cambridge, MA, 1990.

[4] R. D. Beer. Toward the evolution of dynamical neural networks for minimally cognitive behavior. In P. Maes, M. Mataric, J-A. Meyer, J. Pollack, H. Roitblat, and S. Wilson, editors, *From Animals to Animats IV: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pages 421–429. MIT Press-Bradford Books, Cambridge, MA, 1996.

[5] R. Boakes. *From Darwin to behaviourism.* Cambridge University Press, Cambridge, 1984.

[6] A. Cangelosi, D. Parisi, and S. Nolfi. Cell division and migration in a genotype for neural networks. *Network*, 5:497–515, 1994.

[7] Y. Cheneval. Packlib, an interactive environment to develop modular software for data processing. In J. Mira and F. Sandoval, editors, *From Natural to Artificial Neural Computation, IWANN-95*, pages 673–682, Malaga, 1995. Springer Verlag.

[8] A. Clark. *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing.* MIT Press, Cambridge, MA, 1989.

[9] D. Cliff, I. Harvey, and P. Husbands. Explorations in evolutionary robotics. *Adaptive Behavior*, 2:73–110, 1993.

[10] D. C. Dennett. *Consciousness Explained.* Little, Brown and Company, USA, 1991.

[11] M. Dorigo. Editorial introduction to the special issue on learning autonomous robots. *IEEE Transactions on Systems, Man and Cybernetics-Part B*, 26:361–364, 1993.

[12] M. Dorigo and U. Schnepf. Genetic-based machine learning and behavior based robotics: a new synthesis. *IEEE Transactions on Systems, Man and Cybernetics*, 23:141–154, 1993.

[13] G. M. Edelman. *The Remembered Present: A Biological Theory of Consciousness.* Basic Books, New York, 1989.

[14] G. M. Edelman. *Bright Air, Brilliant Fire. On the Matter of the Mind.* Basic Books, New York, 1992.

[15] D. Floreano. Robogen: A software package for evolutionary control systems. Release 1.1. Technical report LabTeCo No. 93-01, Cognitive Technology Laboratory, AREA Science Park, Trieste, Italy, 1993.

[16] D. Floreano. Evolutionary re-adaptation of neurocontrollers in changing environments. In P. Sincak, editor, *Proceedings of the Conference Intelligent Technologies*, volume II, pages 9–20. Efa Press, Kosice, Slovakia, 1996.

[17] D. Floreano and F. Mondada. Automatic Creation of an Autonomous Agent: Genetic Evolution of a Neural-Network Driven Robot. In D. Cliff, P. Husbands, J. Meyer, and S. W. Wilson, editors, *From Animals to Animats III: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, pages 402–410. MIT Press-Bradford Books, Cambridge, MA, 1994.

[18] D. Floreano and F. Mondada. Evolution of homing navigation in a real mobile robot. *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, 26:396–407, 1996.

[19] D. Floreano and F. Mondada. Evolution of plastic neurocontrollers for situated agents. In P. Maes, M. Mataric, J-A. Meyer, J. Pollack, H. Roitblat, and S. Wilson, editors, *From Animals to Animats IV: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pages 402–410. MIT Press-Bradford Books, Cambridge, MA, 1996.

[20] P. Gaussier. Special Issue on Animat Approach to Control Autonomous Robots interacting with an unknown world. *Robotics and Autonomous Systems*, 16, 1995.

[21] D. E. Goldberg. *Genetic algorithms in search, optimization and machine learning.* Addison-Wesley, Redwood City, CA, 1989.

[22] I. Harvey. Species Adaptation Genetic Algorithms: A basis for a continuing SAGA. In F. J. Varela and P. Bourgine, editors, *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, pages 346–354. MIT Press-Bradford Books, Cambridge, MA, 1992.

[23] I. Harvey, P. Husbands, and D. Cliff. Issues in Evolutionary Robotics. In J. Meyer, H. L. Roitblat, and S. W. Wilson, editors, *From Animals to Animats II: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*, pages 364–373. MIT Press-Bradford Books, Cambridge, MA, 1993.

[24] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the theory of neural computation.* Addison-Wesley, Redwood City, CA, 1991.

[25] T. H. Huxley. On the hypothesis that animals are automata. *Nature*, 10:362, 1874.

[26] H. Kitano. Designing neural networks using genetic algorithms with graph generation system. *Complex Systems*, 4:461–476, 1990.

[27] J. R. Koza. *Genetic programming: On the programming of computers by means of natural selection.* MIT Press, Cambridge, MA, 1992.

[28] C. G. Langton. Artificial life. In C.G. Langton, editor, *Artificial Life.* Addison-Wesley: series of the Santa Fe Institute Studies in the Sciences of Complexities, Redwood City, CA, 1990.

[29] W-P. Lee, J. Hallam, and H. Lund. A hybrid GP/GA approach for co-evolving controllers and robot bodies to achieve fitness-specified tasks. In *Proceedings of IEEE 3rd International Conference on Evolutionary Computation.* IEEE Press, 1996.

[30] M. Mataric and D. Cliff. Challenges in Evolving Controllers for Physical Robots. *Robotics and Autonomous Systems*, 19(1):67–83, 1996.

[31] D. J. McFarland. Autonomy and self-sufficiency in robots. AI-Memo 92-03, Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Belgium, 1992.

[32] D. J. McFarland and T. Boesser. *Intelligent Behavior in Animals and Robots.* MIT Press/Bradford Books, Cambridge, MA, 1993.

[33] O. Miglino, H. H. Lund, and S. Nolfi. Evolving Mobile Robots in Simulated and Real Environments. *Artificial Life*, 2:417–434, 1996.

[34] F. Mondada and D. Floreano. Evolution of neural control structures: some experiments on mobile robots. *Robotics and Autonomous Systems*, 16:183–195, 1995.

[35] F. Mondada, E. Franzi, and P. Ienne. Mobile robot miniaturization: A tool for investigation in control algorithms. In T. Yoshikawa and F. Miyazaki, editors, *Proceedings of the Third International Symposium on Experimental Robotics*, pages 501–513, Tokyo, 1993. Springer Verlag.

[36] D. Montana and L. Davis. Training feed forward neural networks using genetic algorithms. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 529–538, San Mateo, CA, 1989. Morgan Kaufmann.

[37] S. Nolfi, D. Floreano, O. Miglino, and F. Mondada. How to evolve autonomous robots: Different approaches in evolutionary robotics. In R. Brooks and P. Maes, editors, *Proceedings of the Fourth Workshop on Artificial Life*, pages 190–197, Boston, MA, 1994. MIT Press.

[38] D. Parisi, F. Cecconi, and S. Nolfi. Econets: Neural networks that learn in an environment. *Network*, 1:149–168, 1990.

[39] R. Penrose. *Shadows of the Mind.* Oxford University Press, Oxford, 1994.

[40] R. Pfeifer and P. F. M. J. Verschure. Designing efficiently navigating non-goal directed robots. In J. Meyer, H. L. Roitblat, and S. W. Wilson, editors, *From Animals to Animats II: Proceedings of the Second International Conference on Simulation of Adaptive Behavior.* MIT Press-Bradford Books, Cambridge, MA, 1993.

[41] D. A. Pomerleau. *Neural Network Perception for Mobile Robot Guidance.* Kluwer Academic Publishing, Boston, 1993.

[42] M. Rudnick. A Bibliography of the Intersection of Genetic Search and Artificial Neural Networks. Technical Report CS/E 90-001, Department of Computer Science and Engineering, Oregon Graduate Center, January 1990.

[43] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Representations by Back-Propagation of Errors. *Nature*, 323:533–536, 1986.

[44] L. Steels. Building agents out of autonomous behavior systems. In L. Steels and R. Brooks, editors, *The "artificial life" route to "artificial intelligence". Building situated embodied agents*, pages 102–137. Lawrence Erlbaum, New Haven, 1993.

[45] B. Yamauchi and R. D. Beer. Sequential behavior and learning in evolved dynamical neural networks. *Adaptive Behavior*, 2:219–246, 1995.

[46] X. Yao. A review of evolutionary artificial neural networks. *International Journal of Intelligent Systems*, 4:203–222, 1993.