Representations in Dynamical Embodied Agents
Re-Analyzing a Minimally Cognitive Model Agent

Marco Mirolli

Istituto di Scienze e Tecnologie della Cognizione,
Consiglio Nazionale delle Ricerche (LARAL-ISTC-CNR),
Via San Martino della Battaglia 44, I-00185 Roma, Italy
marco.mirolli@istc.cnr.it

Abstract. Understanding the role (if any) that the concept of 'representation' might play in cognitive science is a fundamental problem facing the emerging framework of embodied, situated, dynamical cognition. To make progress in this endeavor, here I follow the approach proposed by one of the most influential representational skeptics, Randall Beer: instead of arguing only on the basis of intuitions and speculations, building artificial models of agents capable of minimally cognitive behaviors, and then assessing whether their internal states can be considered to involve 'representations' in any meaningful sense. In this spirit, I operationalize the concept of representing as 'standing in' and I look for the presence of representations in embodied agents which have to solve simple categorization tasks. In a first experiment, no representations can be found, supporting representational skepticism, but the fact that agents which do not possess internal states reach high performance casts doubt on the relevance of the task for the issue of representations. A simple modification of the task makes it more 'representationally-hungry', and in this case successful agents are found to possess internal states that do qualify as representations. I conclude the paper by discussing the benefits of reconciling the embodied-dynamical approach to cognition with the notion of representation.

1 Introduction

In the last quarter century the basic assumptions of Cognitive Science have been questioned by several perspectives: the various approaches of connectionism (Rumelhart et al., 1986b), new robotics (Brooks, 1991), artificial life (Parisi et al., 1990; Steels and Brooks, 1994), adaptive behavior research (Meyer and Wilson, 1990), and dynamical systems (Port and van Gelder, 1995; van Gelder, 1998) have questioned in several ways the classical conception of cognition as the manipulation of symbols according to formal rules. The result is that nowadays a new framework is emerging, which we can call 'Embodied Cognitive Science'. According to this new framework the mechanisms explaining cognition are non-symbolic or sub-symbolic, and cognition itself consists in the adaptation of an agent to its environment, which in turn crucially depends on the interaction dynamics between the agent and the environment it lives in, which can include artifacts and other agents (Bechtel et al., 1998; Pfeifer and Scheier, 1999).

Several of the proponents of this new approach have not only questioned the idea of cognition as symbol manipulation, but the concept of 'representation' itself (e.g. Brooks, 1991; Thelen and Smith, 1994; Wheeler, 1994; Kelso, 1995; Beer, 1995; van Gelder, 1995; Harvey, 1996; Cliff and Noble, 1997; Keijzer, 2001). On the other hand, many other researchers working within the embodied cognition framework keep using a representational language (e.g. Scheier et al., 1998; Nolfi and Marocco, 2001; Sugita and Tani, 2005; Gigliotta and Nolfi, 2008). And several explicit defenses of the concept of representations have been proposed not only by 'traditional' cognitive scientists, but also by active supporters of the embodied cognition framework (e.g. Clark and Toribio, 1994; Clark, 1997b; Bechtel, 1998; Clark and Grush, 1999; Markman and Dietrich, 2000a,b; Prinz and Barsalou, 2000; Spencer and Schoner, 2003; Steels, 2003a). Indeed, the debate about the utility of the concept of representation for cognitive science is ongoing (e.g. Chemero, 2001; Dreyfus, 2002; Haselager et al., 2003; Dietrich and Markman, 2003; Wheeler, 2005; Ramsey, 2007; Garzon, 2008; Gallagher, 2008).

Recently, one of the most influential proponents of both the dynamical approach to cognition and of the skepticism on the notion of representation, Randal Beer (Beer, 1995, 2000), has rightly noticed that

the debate on representations has been mostly based on philosophical argumentations, intuitions, and analogies, rather than on concrete models (Beer, 1996). To overcome this problem, Beer proposed using simple evolutionary simulations of adaptive behavior as a means to ground such a debate on concrete models relying on as few a-priori assumptions as possible. Beer's idea was to construct simple idealized models of 'minimally cognitive behavior' as "prosthetically controlled thought experiments" (Dennett, 1994; Bedau, 2002) with which to explore the fundamental issues posed by the embodied cognition framework: in particular, the issue of representations.

In line with this view, in a recent seminal paper (Beer, 2003b), Beer developed a detailed analysis of an artificial agent capable of exhibiting categorical perception behavior using exclusively methods and concepts of dynamical systems theory. The provision of a detailed explanation of a minimally cognitive behavior in dynamical terms and without relying on the concept of representation is rightly considered by Beer as a substantial contribution in favor of representational skepticism. But notwithstanding the importance of his work, Beer did not bother to check whether in fact his evolved agent's internal states could be considered as representations. In this paper, I try to complete Beer's work: i.e. I assess whether or not (and, eventually, in which conditions) minimally cognitive agents actually use representations. [1]

The rest of the paper is structured as follows. In the next section I briefly discuss the shortcomings of Beer's representational skepticism. In section 2 I describe my replication of the simulations of Beer (2003b). In section 4 I operationalize the basic definition of representations described above in the context of the computational model under investigation and I assess the presence of representations in my best evolved agent, with negative results. In section 5, I present a control simulation with which I check the relevance of the chosen task for the representational debate, again, with negative results. In section 6 I make a simple modification of the task to make it more relevant, and I assess the presence of representations in the best evolved agent in this task, this time with positive results. Finally, in section 7 I discuss the relevance of the presented work with respect to the debate on representations and in section 8 I draw my conclusions.

2 Minimally-cognitive agents and representations

The dynamical analyses of a minimally-cognitive agent exhibiting categorization behavior presented in Beer (2003b) are truly impressive: a brilliant example of the application of the dynamical system approach to cognition. Nonetheless, some perplexities remain about the conclusions that Beers wants to draw on the issue of representations. Several of the commentators of Beer's target article questioned the relevance of Beer's work on the ground that the analyzed model was too simple, and that the same kind of analysis might not scale up to more complex agents (e.g. Edelman, 2003; Clark, 2003). While there might be some truth in this kind of criticism, as I will show later, just *asserting* that more complex cognitive skills *need* to be supported by representations is simply begging the question. The need for representations, as Beer rightly argues in his response (Beer, 2003a), is something which must be demonstrated, not established a-priori.

But even if one accepts to play Beer's game, an even more compelling perplexity arises: do Beer's evolved agents really not possess any representation or have representations simply not been looked for? In fact, though Beer's declared strategy was to "evolve agents on tasks that are rich enough to be representationally interesting, then *examine whether or not these agents actually use representations* in their operation" (Beer, 2003a, pag. 304, emphasis added), Beer's analyses did not include such a careful examination. While a lot of effort was made in providing an impressive amount of dynamical analyses, Beer never explicitly and concretely tried to assess whether representational content might be ascribed to his agent's internal states. Though apparently odd, this is quite typical from researchers in the adaptive behavior community which are critics of representations. Most probably, this is due to the fact that usually these critics consider the notion of internal representation too vague and ill-defined to support such an investigation (Harvey, 1996). Beer seems to share this concern when he says that "'representation' is far too malleable a label" (Beer, 2003b, pag. 237) and that "...this methodology will only work if advocates of particular sorts of representation are willing to specify clear criteria for identifying them" (Beer, 2003a, pag. 304).

Notwithstanding the perplexities of the representational skeptics, close scrutiny of the debate over representations reveals not only that definitions of representation have indeed been given, but also that there is a significant amount of consensus with respect to the basic idea. In fact, all the participants in the debate agree that not all internal states are representations (one of the skeptic's most pressing concerns), and that the simple fact that an internal state *correlates* with some external feature[2] is not sufficient either. Rather, for an internal state to be a representation the correlation must be *functional* to the agent's behavior: in other words, *the internal state must contribute to the agent's behavior in a way that is adaptive to what it correlates with.* For example, Clark claims: "...standing-in requires *not mere correlation*, but adaptively or *functionally intended correlation*" (Clark, 1997b, pag. 464, emphasis added). In the same spirit, Bechtel says: "*...reliable covariation is not sufficient...* the additional component is that a representation has as its *function* the *carrying of specific information*" (Bechtel, 1998, pag. 298, emphasis added). Surprisingly, Beer himself seems to be aware of this consensus as he states: "The basic idea is that these internal representations stand in for (re-present) external things in a way that both *carries a coherent semantic interpretation* and plays a *direct causal role* in the cognitive machinery" (Beer, 2003b, pag. 237). In my view, all these characterizations clearly convey the same basic idea, and are also detailed enough to make the search of internal representations in an simple model agent feasible. Hence, it is time to complete what Beer has left undone: i.e. assess whether minimally cognitive agents actually use internal representations[3] or not.

3 Methods

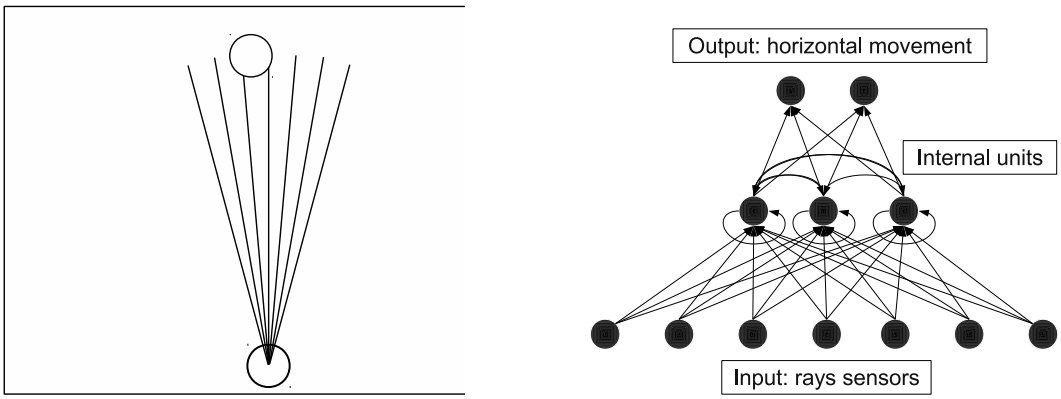
I replicate the categorization experiments presented in Beer (1996, 2003b), with some modifications to irrelevant details (regarding the network's constraints and the genetic algorithm). The agent has a circular body with a diameter of 30 and is situated in an environment of size 400 x 275 (Fig. 1a). The agent's 'eye' consists of 7 rays of maximum length 220 uniformly distributed over a visual angle of $\pi/6$. The intersection between a ray and an object causes an input to be injected into the corresponding sensory node, with a magnitude inversely proportional to the distance to the object. When the rays are at their maximum length, no input is injected, while the maximum input (1) is injected for rays of zero length. The agent can move horizontally as objects fall from above with horizontal velocity proportional to the sum of the opposing forces produced by two motors (maximum speed = 5). The behavior of the agent is controlled by a network with 7 input units (ray sensors) sending connections to 3 fully-recurrent internal units which in turn send connections to the 2 motor units (Fig. 1b). All neurons, including the input ones, are leaky integrators: their activation is updated according to the following equation:

$$o_i = \tau_i o_i^{(t-1)} + (1-\tau_i)(1+e^{-(\theta_i+\sum w_{ij}o_j)})^{-1} \quad (1)$$

where τ_i is a time constant, θ_i is the bias (or the input injected to the sensor in the case of input neurons), and w_{ij} is the weight from neuron j to neuron i.

The parameters of the neural controllers are evolved using a standard genetic algorithm. All the weights, biases and time constants are encoded in the genotype of evolving individuals as 8-bits strings. Weights and biases are normalized in $[-5, 5]$, time constants are normalized in $[0, 1]$. Agents are evolved for their ability to catch circles and avoid diamonds. Each individual is tested for 50 trials, half of which include a circle and half a diamond. In each trial, the agent starts from the middle of the environment and an object falls straight down with a random initial horizontal offset in the range $[-50, 50]$ (with respect to the agent) and a vertical velocity of 3. Circular objects have a diameter of 30 and diamonds have a side of 30.

The fitness of individuals is computed as $f(x) = \frac{\sum p_i}{N}$, where N is the total number of trials (50) and $p_i = 1 - d_i$ for a circular object and $p_i = d_i$ for diamonds; d_i is the horizontal distance between the centers of the object and the agent when their vertical separation goes to 0 in trial i, and is clipped to $maxDist$ and normalized in $[0, 1]$; $maxDist$ is 1.5 times the sum of the radii of the object and the agent (during post-evolutionary tests, I also calculate a discrete trial performance, pd_i, which equals 1 if $p_i > 0.5$, and 0 otherwise). The population is composed of 100 individuals. The 20 best genotypes of each generation are allowed to reproduce by generating 5 copies each, with 3% of their bits re-placed with



(a) (b)

Fig. 1. Experiment's set-up. (a) The agent (gray circle at the bottom) move horizontally while objects (black circle on the top) fall from above. The agent has 7 distal sensors (black lines). (B) The architecture of the neural network.

a new randomly selected value (with elitism: i.e. in each generation the best individual of the previous generation is retained without mutations). Evolution lasts 250 generations.

The basic results of the target experiments were successfully replicated: over 10 evolutionary runs the best evolved agent reached a fitness of 0.999 on the 50 evaluation trials, and a mean performance of 0.9936 when tested on 1000 new randomly generated trials (discrete average performance = 1). This is an almost perfect performance, which is even slightly higher than that of Beer's best evolved agent, which reached an average of 0.9708 on 10000 test trials. Since the focus of the present paper is on searching for internal representations, I will not provide a detailed analysis of agent's behavior. Suffice it to say that, behaviorally, the strategy of our best evolved agent looks qualitatively very similar to that of the agent analyzed in Beer (2003b), being based on an initial foveation of the falling object, an active scanning of it through back and forth movements, and then a decrease of the amplitude of the scanning until circles are centered or a large avoidance movement with respect to diamonds.

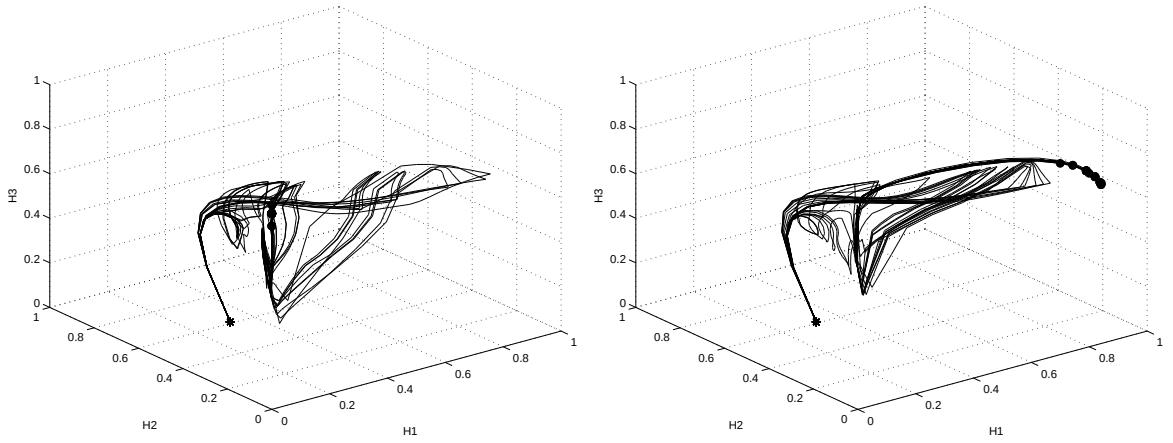
4 In search of the Representational Grail

Now that we have a embodied agent that is able to categorize objects thanks to its dynamical interactions with the environment, we want to assess whether the agent's internal states can be characterized as representations. In order to do that we must operationalize the basic, well accepted notion of representation that we have discussed in section 2. According to such a notion, in order to qualify as representational an internal state must: (a) tend to correlate with some relevant (for the agent) feature of the environment and (b) contribute to agent's behavior in a way that is functional (adaptive) to the feature with which it correlates. Fortunately, the simplicity of the experimental set-up makes the task of operationalizing these two requisites relatively easy.

Since all the agent has to do is catching circles and avoiding diamonds, and since the agent's simple neural network cannot do very complex perceptual computations, the most plausible (if not the only) features that our agent might represent are just circles and diamonds. Hence, for checking criterion (a) we must look at whether any of the agent's internal states do indeed reliably correlate with circles and diamonds. The agent control system is composed by the input, hidden, and output units and by the connection weights (linking input-to-hidden, hidden-to-hidden, and hidden-to-output). The inputs constitute the agent's sensory system that is directly activated by the objects in the environment, and the outputs constitute the agent's motor system that directly affects behavior. Hence, inputs and outputs constitute the interface between the system and the environment and cannot be considered part of its internal machinery. The connection weights are what determine the behavior of the system given the input received from the environment. Since they never change during the whole 'life' of the agent, they

do not correlate with anything, and hence they cannot represent anything. The only things that are internal to the network and that do change during the agent interactions with the environment are the activations of the hidden units. Hence, for checking criterion (a) we only need to look at the activations of the 3 hidden units and check whether they do correlate or not with circles and diamonds (since the pattern of the activations of the 3 hidden units is the only thing that is internal to the agent and that changes through time, I will refer to that pattern in each point in time as the 'internal state' of the agent at that point in time)[4].

Fig. 2 shows the trajectories of the internal state of the best evolved agent during 50 tests, half of which involve interacting with circles and the other half with diamonds. While individual trajectories during different interactions with the same kind of object vary depending on the object's initial offset, the agent's internal state at the end of the trial does correlate very well with the category of the object with which the agent has interacted (and with the action that the agent has performed: catch or avoid). In fact, the vector of the activations of the hidden units is very similar every time the agent caught a circle (the centroid of end points of the trajectories with circles is $< 0.0000, 0.0002, 0.8741 >$) and very similar every time the agent avoided a diamond (in this case, the centroid's coordinates are $< 0.8190, 0.0186, 0.7542 >$); but the centroids in the two cases are quite different, in particular with respect to the activation of the first hidden unit.



(a) Circles (b) Diamonds

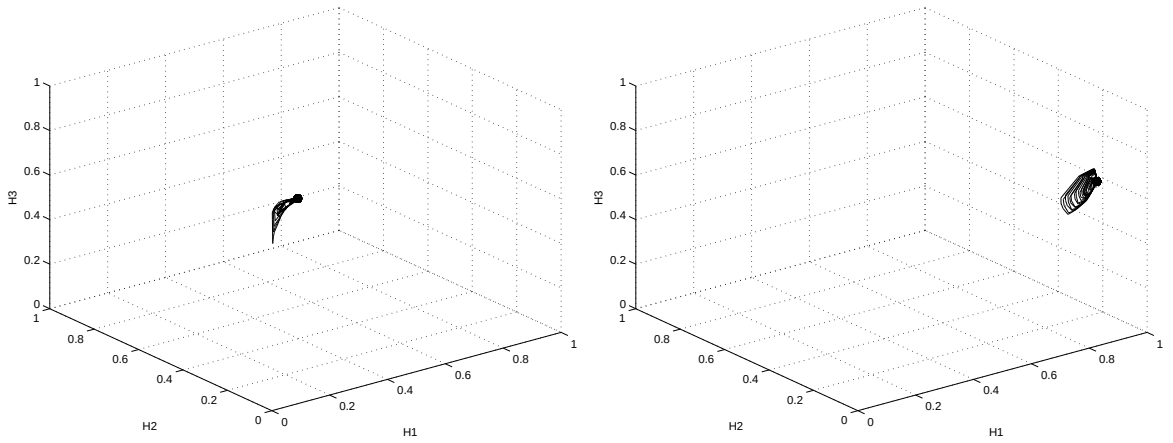
Fig. 2. Correlation test for the categorization experiment. Graphs report the trajectories of agent's internal state during trials with circles (a) and diamonds (b). Each dimension corresponds to the activation of one of the three hidden units. Stars represent the starting point (equal for all the trials), while filled circles represent end points

Now that we have discovered a significant correlation between the patterns of activation of the hidden units at the end of the trial and the category to which objects in the environment belong, we have a plausible candidate for a representation: that is, the centroid of the trajectories' end-points for each category. In order to check whether these internal states can be considered real representations we must ensure that this correlation is functionally significant, as required by criterion (b) above. How can criterion (b) be operationalized in a simple embodied system like our own? Since the requirement is that the candidate representation must contribute to the agent's behavior in a way that is adaptive with respect to what it is supposed to represent, the most straightforward way to operationalize this criterion is to *impose* the candidate representation in the agent's 'brain' and check whether the resulting behavior is functional with respect to what is supposed to be represented. In particular, I implement this by imposing the candidate representations *before* the interaction with the object has started (i.e. at the beginning of the trial). If the resulting behavior is (immediately) adaptive with respect to the supposedly represented feature (in our case object's category), than the criterion is met, and we can say to have found 'real' representations; otherwise, the criterion is not met, and we have discovered that our

agent does not represent anything after all, as Beer suggested[5]. I run 50 random tests (half with circles and half with diamonds) in which I initialize the activations of the internal neurons of the agent to the centroid of the final patterns of the category of the object with which the agent has to interact (in this way, the agent is provided with the candidate representation of the correct category a few time steps –about 10– before the falling object can impact the sensors). All other units are initialized, as usual, to 0.

The results of such a test are quite interesting. Imposing the candidate representations onto the agent before it can interact with the object does not facilitate the agent. On the contrary, average performance is disrupted to chance level: 0.5 for both the continuous and the discrete measure. In fact, though starting with different internal states, agent's behavior is almost identical in both circles and diamond trials: the agent always move right, thus always producing an avoidance behavior irrespective of the category of the falling object (thus scoring 1 with all the diamonds and 0 with all the circles). The reason is that even though the regions in the internal state space correlating with the two categories of objects seem to correspond to two attractors, from which the internal dynamics tend not to move (Fig. 3), the attractor correlating with circles does not, by itself, make the agent behave in a way which is appropriate to circles, i.e. staying still while centering the falling circle. Rather, it is only when the internal state is coupled with the appropriate agent-environment dynamical interactions that the whole system composed by the state, the agent, and the environment produces the appropriate behavior. Hence, the states correlating with the presence of circles cannot be considered a representation of circles, since they satisfy only one (the first) of the two criteria for representationhood.

At this point, a strenuous defender of representations might claim that even if we have not found a representation of 'circles' at least we have found a representation of 'diamonds'. This would be misleading because there is substantial agreement in the literature that for an internal state to be a representation it must be part of a *system of representations* in which different states represent different things. If this is not the case, then the use of the notion of representation is of no value. Hence, we are forced to conclude that the internal states of our model do not have representational value. In fact, we cannot ascribe to those internal states any coherent semantic interpretation since they cannot appropriately guide behavior independently from their appropriate coupling with the agent-environment interactions.



(a) Circles (b) Diamonds

Fig. 3. Representation test for the categorization experiment. Graphs report the trajectories of agent's internal state during trials with circles (a) and diamonds (b). In this test, the starting point for circles is the candidate for the 'circle' representation (at $< 0.0000, 0.0002, 0.8741 >$), while the starting point for diamonds is the candidate for the 'diamond' representation (at $< 0.8190, 0.0186, 0.7542 >$)

5 Representational 'Hungryness'

Now that we have assessed that no representational role can be ascribed to our minimally cognitive agent, a doubt can still come to mind (acknowledged by Beer himself): is the chosen task sufficiently representationally-hungry (Clark and Toribio, 1994) to "fully engage questions regarding the utility of representational thinking" (Beer, 2003b, pag. 239)? Or is it too simple to be relevant for the debate, as some of Beer's critics (e.g. Edelman, 2003; Clark, 2003) suggest? With respect to this important issue Beer seems to betray his own methodological rigor. In fact, he addresses the issue at the same level of his critics, i.e. with verbal argumentations and by appealing to intuitions (Beer, 2003b,a), instead of trying to ground the debate on concrete modeling work. In fact, even though we currently have no way to measure 'representational hungryness', at least we can set a clear, operational threshold for what cannot be said to be representationally hungry in any meaningful sense: *a task is surely not representationally hungry if it can be solved by agents which do not possess internal states.*

In the spirit of the minimally cognitive behavior approach, with our operationalized threshold we can start to ground the debate on (minimal) representational hungryness on concrete simulations rather than on intuitions and verbal argumentations. In order to do this, I re-ran the same evolutionary experiments described above with the only difference that in this case agents' neural network has no hidden units, and the 7 input units are directly connected to the 2 output units. Since the only thing that is 'internal' (in between the inputs and the outputs) to such a network is the weight connection matrix, which does not change during the agent's 'life' and hence cannot correlate with anything, the agents of this simulation have no internal states: hence they cannot, *by definition*, have internal representations.

The results of these simulation are quite illuminating: over 10 evolutionary runs the best evolved agent reached a fitness of 0.97 on the 50 evaluation trials. The average performance of this agent on a test with 1000 new randomly generated trials was 0.9027, with a discrete average performance of 0.93. This means that the agent mis-categorize the object only in 7% of the trials. All errors happen with circles, in particular when they start to fall at the very left with respect to the agent. In those cases, the agent simply ignores the object: it moves right from the very beginning, immediately loose 'visual' contact with the object, and ends the trial with an erroneous 'avoid'. Every time the agent interacts with the object, it correctly categorizes it, catching circles and avoiding diamonds.

From these results, it is clear that the task cannot really be considered suitable to assess the utility of representational talk, as it can be solved by an agent that has no internal states and, *a fortiori*, cannot have any internal representations. Note that this does not mean that we have to raise the threshold for considering a behavior cognitively interesting so that categorical perception does not qualify. Categorization is surely one of the most (if not *the* most: Harnad, 2005) fundamental cognitive phenomena, and since our agents are able to produce a minimal form of categorization they do qualify as minimally cognitive. But if agents which *cannot* possess representations are able to solve the task, this means that the task, though relevant for cognitive science, is not relevant to the representational issue. In fact, this is indeed another confirmation of one of the main contributions that embodied cognition research has provided to cognitive science: namely, the realization that at least *some* of the problems which, from a classical stance, might seem to require the use of internal representations can in fact be solved by relying on pure sensory-motor, non representational strategies (Brooks, 1991; Nolfi, 1998; Pfeifer and Scheier, 1999). As discussed above, to be relevant not only for cognitive science in general, but for the debate on representations in particular, a task should at least be such that it cannot be solved by agents without internal states. Again, this does not mean that we have to stop discussing the issue until we will manage to evolve agents that are able to perform extremely complicated tasks requiring, for example, abstract thinking and problem solving abilities or the use of language, as Beer's critics seemed to suggest. Even though it is always difficult to establish a priori whether a task is solvable or not by purely sensory-motor strategies, we can use our evolutionary simulations to find this out 'empirically'.

To do so, we just need to find the simplest modification of the task we have used so far that makes it pass the operational threshold we have just set for representational hungryness: i.e. we need to modify it so that it cannot be solved by agents that have no internal states. One possible way of achieving this might consist in making the task require some (minimal) form of memory: if the task requires that the agent 'remember' the object category also after it has stopped its dynamical interactions with the object, it seems that an agent without internal states cannot solve it. A task with similar characteristics can

be obtained by just slightly changing the behavior that is required from the agent: instead of avoiding diamonds and catching circles, you just need to ask the agent to move as leftward as possible with circles and as rightward as possible with diamonds. In this way, once the agent has categorized the object and moves towards the appropriate direction, the object disappears from sight. In such a situation, in order for the agent to keep on moving correctly it seems to require some form of memory about which kind of object it has interacted with.

In order to 'empirically' test whether such a simple modification would indeed be sufficient for making the task more representationally-hungry (according to our operational threshold), I ran another set of simulations in which the agents' neural networks have no internal units and the categorization task is changed accordingly. In particular, in this new experiment everything runs as in the previous simulation except for the fitness function. In this case, the contribution to fitness in each trial i is computed as follows: $p_i = 1 - d_i$ for a circular object and $p_i = d_i$ for diamonds, where d_i is the normalized distance between the agent and the left wall of the environment when the vertical separation between the agent and the falling object goes to 0[6].

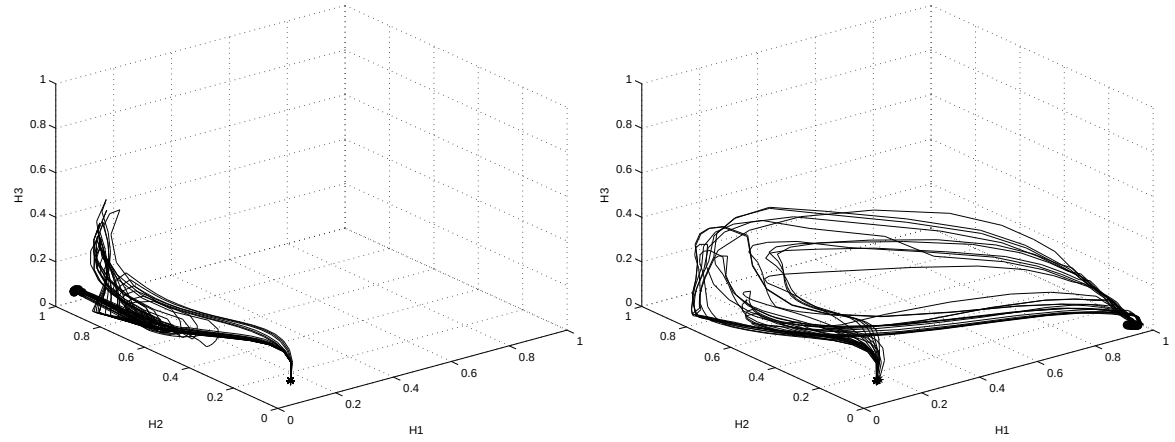
The results of the simulations confirm our analysis. Indeed, purely reactive agents are not able to evolve a successful strategy: in this case over 10 evolutionary runs the best evolved agent reached a fitness of 0.675 on the 50 evaluation trials, and a mean performance of 0.5985 on a test on 1000 new randomly generated trials. Agent's behavior has little to do with categorization: it tends always to move leftward, while, sometimes it centers the object when it falls from the left. As a result, the discrete average performance of this agent is 0.56, meaning that the ability of the agent to categorize objects is just very slightly higher than chance level.

6 Representations, at last

The slightly modified task just described is considerably more relevant for a discussion about representation than the original one, as it seems to require the possession of internal states to be solved (of course, it is possible that other kinds of neural networks without hidden states but, for example, with a memory of the motor states might be able to solve the task; nonetheless, if we do not provide this kind of information to the network, the previous simulation shows that this task requires internal states). So, I ran a final set of evolutionary experiments in which the neural network is that of the original task (with three fully recurrent internal units) but the task is the slightly modified one, i.e. going to the left for circles and to the right for diamonds. Agents provided with internal states are able to efficiently solve the task: over 10 evolutionary runs the best evolved agent reached a fitness of 0.923 on the 50 evaluation trials, and a mean performance of 0.8985 on 1000 new randomly generated trials (discrete average performance = 1). Now we can apply the two tests for assessing the presence of representations to the best evolved agent in this task. The results are shown in Fig. 4 and 5. As in the original experiment, the internal state of the agent always ends up in one of two very small regions of its state space depending on the object the agent has interacted with (Fig. 4). But, contrary to what happened to the agent of the original experiment, in this case the internal state correlating with the object's category consistently guides the agent's behavior in a way that is appropriate to what it correlates with (i.e. the category).

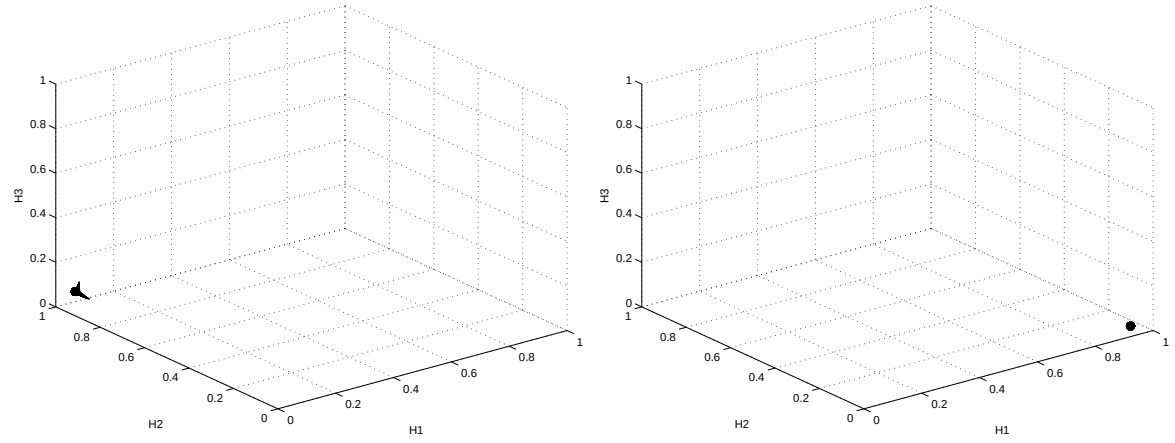
If we impose this internal state before the agent has interacted with the object (i.e. at the beginning of a trial), the internal state is maintained constant during the whole trial (Fig. 5)[7]. As a result, the agent always produces the appropriate behavior: moving leftward in case of circles and rightward in case of diamonds. Indeed, the performance in these tests is even significantly better (0.9568) than in the normal tests (0.8985). The reason is clear: since in this case the agent 'knows' from the beginning of the trial which object is present, it can directly produce the appropriate behavior without wasting time in actively scanning the object in order to discover its category, as it needs to do in normal trials. Hence, the two attractors in the agent's internal state space satisfy both the criteria for representationhood: they both (a) correlate with a feature in the environment that is relevant for the agent's survival and reproduction (i.e. the category of falling object), and (b) guide agent's behavior in a way which is adaptive with respect to that feature (i.e. moving leftward or rightward).

Having discovered that our agent solves its task by relying on internal representations we can use this knowledge to make predictions about the agent's behavior. For example, we can predict that if the agent



(a) Circles (b) Diamonds

Fig. 4. Correlation test for the modified experiment



(a) Circles (b) Diamonds

Fig. 5. Representation test for the modified experiment

'thinks' that a circle is present in its environment (i.e. if its internal state is the one representing circles) it will behave accordingly, irrespectively of the effective presence of a circle. I tested this prediction by running another set of 50 trials in which I impose, at the beginning of the trial, the internal state representing the *wrong* object. The prediction was verified: in the 100% of cases the agent behaved in a way which was appropriate to the category of the (wrongly) represented object, and not to the one actually present in the environment. As a result, the resulting average performance was 0.0448 (discrete average performance = 0).

7 Discussion

Understanding the role (if any) that the concept of representation might play in the future science of behavior is one of the most important, fundamental problems facing the emerging framework of embodied cognition. To make progress in this endeavor, in the present article I have followed the approach recently proposed by one of the most influential representational skeptics, that is Randall Beer (Beer, 1996, 2000, 2003b): instead of continuing to argue endlessly on the base of intuitions and speculations, build artificial, idealized models of agents capable of minimally cognitive behaviors, and then assess

whether their internal states can be meaningfully considered to involve 'representations'. In this spirit, I looked for the presence of representations in evolved agents which had to solve Beer's well-known categorical perception experiment. In order to do that, I had to operationalized a fairly standard concept of representations (Clark, 1997b; Bechtel, 1998; Beer, 2003b) according to which an internal state can be considered a representation if (a) it tends to correlate with some relevant (for the agent) feature of the environment and (b) it contributes to agent's behavior in a way which is functional to the feature with which it correlates. While I was able to find states satisfying (a), criterion (b) was not met, thus supporting Beer's view that no representations could be found in these agents. On the other hand, the successful replication of the same experiments with agents whose controller did not possess any internal state and hence could not have internal representations by definition questioned the relevance of the chosen task for a fair assessment of the representational issue. A simple, slight modification of the original task made it unsolvable by agents with no internal states, and consequently much more suitable for our purposes. The assessment of the presence of representations in agents provided with internal states able to solve the modified task gave positive results: in this case, I found internal states that not only correlated with environmental features of the environment, but also guided agent's behavior in a way that was functional with respect to those features, thus meeting both above-mentioned criteria for representationhood.

A few points still need to be clarified. First, the operationalization of the concept of representation I have provided is surely very limited and simple. In fact, I do not consider it as a full-blown definition of representation, but rather as a working characterization with which we can at least ground future research on the topic. Indeed, the most important role that this characterization is supposed to play in the debate on representations is that of preventing any future complaint that the notion of representation is too vague and malleable to be of any use in a rational debate (e.g. Harvey, 1996; Cliff and Noble, 1997; Haselager et al., 2003; Beer, 2003b). My operationalization can no doubt be improved, but it has at least three undeniable merits: (1) it is not vacuous, (2) it is in line with the standard, pre-theoretic notion of representation, and (3) it can be used for assessing the presence (or absence!) of representations in embodied agents.

Second, the notion of internal representation I have discussed in this paper is not only meaningful and applicable, it is also predictive, and hence it can be profitably used for understanding an agent's behavior. For example, in the last simulation I have successfully tested the prediction that the agent would behave according to its internal representation of objects irrespective of whether the object is actually present in the environment. This kind of test demonstrates another feature which is considered to be characteristic of representations, that is the possibility to have *wrong* representations : "it seems an essential aspect of representation that misrepresentation is possible" (Bechtel, 1998, pag. 298; see also, for example, Millikan, 1984; Cummins, 1989).

Third, it is probably worth clarifying the relationships between the broad notion of representation that has been used in this paper and narrower notions. For example, Clark and Grush (1999) distinguish between a weak and a strong sense of representations (see also Clark, 1997b): for a state to classify as a representation in the weak sense, only adaptive functional correlation is required; while representations in the strong sense are "inner items capable of playing their roles in the absence of the on-going perceptual inputs that ordinarily allow us to key our actions to our world." (pag. 9). This stronger notion is in line with the idea that representations are related to the internal tracking of things that are currently absent (Haugeland, 1991; Smith, 1996). Clark and Grush introduced the distinction in order to discriminate between the kind of inner states that might be present in systems like the one I analyzed in the present paper and the inner states of an emulator in the control-theoretic sense: i.e. states that constitute predictions of the perceptual consequences of the system's actions. This notwithstanding, it is interesting to note that in order to make my agent develop internal representations (in the weak sense), I modified the original task so that the agent had to behave appropriately to the object's category *even when not directly perceiving the object*. This prevented the possibility for the agent to rely on a purely sensory-motor reactive strategy: in fact, the category of the perceived object had to be internally stored so that the agent might keep on producing the appropriate behavior after the object had exited the agent's visual field. In other words, the function of my agent's representational states is just to play the role of stand-ins for the category of the perceived object when the object is absent to the agent no longer providing any

perceptual input. Hence, our minimalist simulations seem to demonstrate that it is possible to have an agent possessing "full-blooded" representations (in the strong sense of Clark and Grush) without having coupled inverse and forward models, as the emulator theory of representations suggests (Grush, 2004). To clarify my position: I agree that there are good reasons to believe that something like emulators developed in real animals for motor control and that they also might play a fundamental role in higher level cognitive activities (Barsalou, 1999; Hesslow, 2002; Grush, 2004; Pezzulo, 2008), but the usefulness of the (broader) concept of representation is not limited to the cases in which emulations (or simulations) are present.

Fourth, the previous issue is related to another general point that is worth clarifying. Most accounts of representations involve references to some kind of 'cognizer' to which representations must be delivered in order for their representational status to be considered as effective. This is consistent with the intuitive notion that the representational status of a given entity crucially depends on the 'consumer' of the representation: the same object, action, or event can represent different things (including no-thing) to different agents. It is on this grounds that representational skeptic Inman Harvey stated that "when talking of representation it is imperative to make clear who the users of the representation are" (Harvey, 1996). How does our approach deal with the problem of identifying the user of the representation? The first thing to consider is that if the suggested 'cognizer' is taken too literally (as it is typically done), it just begs the question about representationhood: in its cognitive activities the cognizer would need to manipulate its own representations, thus requiring another sub-cognizer to which its own representations have to be delivered, and so on. If one wants to escape from an infinite regress, there must be a final 'user' of the representation that is simple enough not to contain any representation (the potential problem of infinite regress in cognitive explanations is well-known in the philosophy of mind: for a valuable discussion, see, for example, Dennett, 1978). In this respect, a simple three-layer neural network like the one used in this work can be conceived as being composed of two sub-systems: the first, consisting of the connection weights linking the input to the hidden units plus the recurrent connections in the hidden layer, is the part of the system that 'produces' the representations; the second, consisting of the connection weights linking the hidden to the output units, is the part of the system that 'uses' the representations produced by the other sub-system. This way of framing the problem is interesting because, on the one hand, it satisfies the intuition that representations require a user (and thus responds to Harvey's concerns); on the other, it demonstrates that this requirement does not necessarily lead to an infinite regress of homunculi inside the cognitive system.

Fifth, I think that the exercise of looking for representations in a principled way has proved to be quite instructive and rewarding. For example, it has permitted us to verify once again that there are cognitively interesting tasks, like the original categorization task, that can be solved by relying only on purely sensory-motor reactive processes (for previous demonstration of this point see, for example, Harvey et al., 1994; Nolfi, 1997; Scheier et al., 1998; Floreano et al., 2004). It has also allowed us to discover that other tasks, like the modified task involving memory, are much less tractable with purely sensory-motor strategies (on this point, see also Nolfi, 2002; Mirolli et al., 2010). Furthermore, and most importantly, our exercise has made us discover another interesting fact: agents endowed with internal states that have to solve the first kind of simpler tasks tend to rely on sensory-motor strategies and not develop internal representational states. On the other hand, the results of our last simulations suggest that agents dealing with more complex tasks that can't be solved by relying on purely sensory-motor strategies might indeed tend to develop internal states to which we can ascribe a representational status. However, it is important to remember that on the question which kind of tasks requires which kind of ability we can only provide partial indications of general tendencies: definite conclusions and general rules can seldom, if ever, be drawn. To clarify: my simulations demonstrate that, *in this particular case*, making the task more memory-hungry resulted in making it more representationally hungry. This by no means implies that all and only the tasks that require some form of memory need internal representations to be solved. As discussed above, artificial life / evolutionary robotic research has already sufficiently demonstrated that our intuitions on what is the space of possible solutions to a given task can be completely wrong, and that whether or not a task requires representations (and which kind of representations) cannot in general be established a priori. The only thing we can do is to keep on investigating, both through real experimentation and through simplified computational models, which kinds of tasks *tend to require*

which kinds of abilities. In this respect, our simulations do suggest that memory-hungry tasks tend to be representationally-hungry. But clearly much more work has to be done in order to find out which kinds of task tend to *promote* or even *require* the development of internal representations and which ones do not. And I think that the kind of idealized models that have been used in this paper constitute very appropriate tools for investigating this kind of issues.

Finally, it might be useful to briefly clarify my position with respect to the standard account of representations in Cognitive Science. The view of cognition that is emerging from embodied, situated, dynamical approaches strongly contrasts the old view that cognition can be understood as the manipulation of the kind of objective, discrete, abstract, amodal, rule-governed representations assumed by classical cognitive science. Indeed, even though it is possible that *some* aspects of cognitive activity might involve the use of internal representations that have *some* of the features of the representations assumed by classical cognitive science, as argued, for example, by Markman and Dietrich (2000b,a), I think that the best way to understand high-level human cognition is to investigate how lower-level sub-symbolic cognitive functions are transformed by the internalization of 'cognitive tricks' mediated by language. The idea, originally developed by Vygotsky in the 1930s (Vygotsky, 1978) and recently re-proposed in cognitive-science oriented philosophy of mind (Dennett, 1991, 1993; Clark, 1997b, 2006) and in several areas of psychology (e.g. Gentner, 2003; Spelke, 2003; Tomasello, 2003), is that human cognition depends on the internalization of the social interactions that a developing child has with adults and more skillful peers. The behavior of the child is continually influenced by social interactions, typically mediated by language, that help the child in solving all kinds of (cognitive) tasks: linguistic social aid helps the child in learning how to categorize experiences, to focus attention on important aspects of the environment, to memorize and recall useful information, to inhibit un-appropriate spontaneous behavior, to construct plans for solving complex tasks, and so on. During development the child learns to rehearse the social linguistic aid when faced with similar tasks all alone, and then, when the linguistic self-aid has been mastered, private speech is internalized, thus becoming inner speech. If this analysis is correct, human cognition just *seems to be* constituted by language-like symbol manipulation because it is the result of the transformation of sub-symbolic embodied and dynamic processes by internalized linguistic social aids (for more detailed discussions of these ideas, see Mirolli and Parisi, 2009, 2010; for preliminary modeling works along these lines, see Schyns, 1991; Cangelosi and Harnad, 2000; Steels, 2003b; Clowes and Morse, 2005; Lupyan, 2005; Mirolli and Parisi, 2005a,b, 2006).

8 Conclusion

The main goal of the present paper was to show that the new embodied cognitive science is not in contrast with the idea of internal representations per se, but only with the symbolic representations that were prototypical of classical cognitive explanations. If this is the case, as I hope I have demonstrated, what must be done in order to further develop the new embodied cognitive science is to stop quarreling about the necessity of the concept of representation in general and to focus on the development of new ideas and theories regarding the kinds of representations that embodied systems might need to use for solving their cognitive tasks. Indeed, new ideas should be developed in order to re-think all aspects of representationhood, including:

1. What is represented: not the world as it is but rather possibilities for action, ways of interacting with the environment, and so on. This idea is in line with the recent psychological research on the notion of affordance (Gibson, 1979; Tucker and Ellis, 2001; Caligiore et al., 2010), with the neuroscientific findings related to canonical neurons (Rizzolatti et al., 1988; Jeannerod et al., 1995; Murata et al., 1997; Rizzolatti and Sinigaglia, 2008), and with the notions of 'pushmi-pullyu' or 'action-oriented' representations proposed by several philosophers (e.g. Millikan, 1996; Clark, 1997a).
2. How things are represented: not through language-like, amodal symbols but rather through action-related and/or modality-specific vehicles. Several recent proposals in both psychology and neuroscience support this view: see, for example, the various concepts of 'embodied memories' (Glenberg, 1997), 'perceptual symbol systems' (Barsalou, 1999; Prinz and Barsalou, 2000), 'sensory-motor emulators' (Grush, 2004), and 'mental simulations' (Jeannerod, 2006). An important issue is to investigate

the form that such kinds of representational vehicles may take: patterns of neural activations, trajectories of patterns of activations, attractor states... (see ,e.g. Rumelhart et al., 1986b; Elman, 1990; Spencer and Schoner, 2003; Gigliotta and Nolfi, 2008).
3. How representations are created: not through the passive processing of static perceptions but rather through an active process that strongly relies on the dynamical interactions with the environment. Again. several lines of research point in this direction (see, for example Yarbus, 1967; Ballard, 1991; Churchland et al., 1994; Findlay and Gilchrist, 2003; Noë, 2004; Mirolli et al., 2010).
4. How to study and understand all this: there is indeed an urgent need to develop not only new theories and new concepts, but also new experimental and computational methods for studying and analyzing embodied cognitive behavior.

I endorse that this endeavor will require the fruitful collaboration between open-minded philosophers, psychologists, neuroscientists, and computational modelers.

Notes

[1]It might be also useful to clarify from the very beginning what I *do not* do here. In particular, I do not *argue* for the dynamical, embodied, situated cognition approach to cognitive science. Here, I just take the view which considers cognition as "environmentally embedded, corporeally embodied, and neurally embrained" (van Gelder, 1999, pag. 244) for granted, and contribute to the debate about whether such a view implies a rejection of the notion of representation or not. For detailed discussions in favor of the embodied approach to cognition, see, for example, (Varela et al., 1991; Thelen and Smith, 1994; Port and van Gelder, 1995; Clark, 1997a; Pfeifer and Scheier, 1999; Beer, 2000; Pfeifer and Bongard, 2006). Furthermore, I do not aim at providing a *theory* of representations and representational content, as many others have been doing (e.g. Fodor, 1981; Millikan, 1984; Dretske, 1988; Cummins, 1989; Barsalou, 1999; Bickhard, 1999; Grush, 2004; Pezzulo, 2008). I just take the pre-theoretic, intuitive concept of representation that seems to be shared by most (if not all) authors discussing about representations, I operationalize it, and I use this operationalization to assess whether minimally cognitive situated embodied agents can be meaningfully ascribed representational states.

[2]I will use the term 'feature' to refer to any possible content of a representation: objects, categories, events, actions, contexts... Indeed, one of the most fundamental and challenging issues on representation is to understand what, if anything, is represented.

[3]I what follows, even when I use the term 'representation' without further qualifications I actually mean 'internal representation', where with 'internal' I just mean 'in between the input and output of a system'. This is just to clarify that here we are not interested in external representations like paintings, writings, signs posed in the environment, and so on, nor in cases in which an embodied agent uses "the world as its own best model" (Brooks, 1990). Here we are exclusively dealing with the psychological notion of 'representation' as an internal stand-in for external reality. Hence, an agent that has no internal states, like the one that will be described in section 5, cannot, by definition, have representations in our sense

[4]The idea of looking at the pattern of activations of the hidden units for candidate representations in a neural network is certainly not new: rather, it is the standard approach in connectionist research at least since the seminal paper by Rumelhart et al. 1986a. The only difference of our work with respect to the connectionist literature is that our neural network is embodied, that is it controls the behavior of an agent that interacts with its environment. With respect to the representational issue this can make a big difference.

In a dis-embodied system like a classical connectionist network the 'behavior' of the network corresponds to the single output pattern given by the network in response to a single input pattern. In order to behave correctly the network has to transform the patterns in the input units, which might not correlate appropriately with the required response, in patterns in the hidden units that do correlate with the required output, so that the appropriate output can be provided. Hence, in such dis-embodied systems, one can be almost certain that, as long as the system behaves correctly, internal representation can be found, because the activation patterns in (sub-sets of) the hidden units will certainly (a) correlate with something relevant – i.e. the category of the input that determines what is the appropriate response – and (b) contribute to drive the output of the system in a manner that is appropriate to what they correlate with – because the appropriate response depends just on those patterns.

Taking into account embodiment makes two critical differences with respect to the issue of representations. First, the presence of an internal state that would be classified as a representation cannot be taken for granted any more, because in an embodied system behaviors crucially depend not just on internal states, but rather on the dynamical interactions between the system and its environment. Then, in contrast with dis-embodied neural

networks, in which appropriate responses must depend on 'meaningful' internal states, in embodied systems the dynamical interactions with the environment can, at least in principle, result in adaptive behavior without any internal state correlating with any relevant feature of the environment. This is the ground on which proponents of embodied cognition and/or of the dynamical approach has based their skeptical challenge to the notion of representation. Second, in an embodied system there is no longer a single internal state that corresponds to a single input and to a single response. This is why in order to search for internal representations we need to look at the whole sequence of internal states produced during the agent's interaction with the environment and try to find a sub-set of such states that does correlate with something relevant in the environment.

Finally, the idea of looking for representations in the (dynamics of the) patterns of activations of a neural network's hidden units is also in line with what happens in neuroscience. When looking for internal representations in real brains the vast majority of neuroscientists look at correlations between external events and the frequencies of spikes (firing rate) of neurons, which is just what the activation of an artificial neuron corresponds to. There are other, complementary proposals, according to which some information in the brain might be passed through the timing of single spikes, or through the synchronization between neurons or groups of neurons (see, for example Singer, 1999; VanRullen et al., 2005). Although this is certainly an interesting possibility which merits further investigation, such coding cannot have any correspondent in neural network models in which neurons are firing rate units and do not have spikes. Since our model is of this latter kind, it simply cannot represent through spikes, and candidate representations can only be found in the patterns of activations of the hidden units.

[5]Note that, in line with the intuitive notion of representation that we are trying to operationalize here, this second criterion for representationhood is purely behavioral, and does not make any assumption with respect to what should be happening inside the system. The internal state might be a point attractor that never changes throughout the trial, it might follow a cyclic trajectory, or it might (at least in theory) even change chaotically. As far as our criterion for representationhood is concerned, the only thing that matters is that the internal state leads to an external behavior that is appropriate with respect to the object category that is is supposed to represent (either circle or diamond)

[6]It is important to bear in mind that this modification to the task for increasing its representational hungryness is not the only one possible: it is only the simplest and most straightforward that could do the required job. In particular, though I made the task more representationally-hungry by making it more memory-hungry, this must not be taken to imply that there is a special, necessary relationship between the need of representations and the need of memory. Indeed, there may very well be tasks that do not require memory but that are nonetheless representationally-hungry, and, viceversa, tasks that require some form of memory that do not require representations. Neither our operational criteria for representationhood nor the operational threshold for representational hungryness make any reference to the need of memory. It just happened that, in this particular case, adding a need for memory in the task was sufficient for making it pass the operational threshold for representational hungryness.

[7]The fact that the internal states that do qualify as representations are states that are maintained constant throughout the trial is contingent to this particular case, and must not be considered as necessarily related to representationhood. Our operational criteria for representationhood do not mention any stability requirement. Hence, there may well be internal states that are not enduring but that qualify as representations and, viceversa, it is certainly possible to have enduring internal states that do not possess the requisites for qualifying as representations. For example, even the states that were tested for representationhood in the original task in section 4 happened to be maintained relatively constant during the trial (see figure 3), but they did not pass the test for representationhood.

Acknowledgements

I thank Domenico Parisi, three anonymous reviewers, and the journal Editor for their comments that helped to substantially improve the paper.

Bibliography

Ballard, D. (1991). Animate vision. *Artificial Intelligence*, 48(1):1–27.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–609.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist challenge in cognitive science. *Cognitive Science*, 22(3):295–318.

Bechtel, W., Abrahamsen, A., and Graham, G. (1998). The life of cognitive science. In Bechtel, W. and Graham, G., editors, *A companion to cognitive science*. Blackwell, Oxford, MA.

Bedau, M. (2002). The scientific and philosophical scope of artificial life. *Leonardo*, 35(4):395–400.

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72(1-2):173–215.

Beer, R. D. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. In Maes, P., Mataric, M., Meyer, J., Pollack, J., and Wilson, S., editors, *From animals to animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pages 421–429, Cambridge, MA. MIT Press.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, 4(3):91–99.

Beer, R. D. (2003a). Arches and stones in cognitive architecture: Reply to comments. *Adaptive Behavior*, 11:299–3055.

Beer, R. D. (2003b). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11(4):209–243.

Bickhard, M. H. (1999). Interaction and representation. *Theory and Psychology*, 9(4):435–458.

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, 6:3–15.

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

Caligiore, D., Borghi, A., Parisi, D., and Baldassarre, G. (2010). Tropicals: A computational embodied neuroscience model of compatibility effects. *Psychological Review*, 117:1188–1228.

Cangelosi, A. and Harnad, S. (2000). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, 4:117–142.

Chemero, A. (2001). Dynamical explanation and mental representations. *Trends in Cognitive Sciences*, 5(4):141–142.

Churchland, P., Ramachandran, V., and Sejnowski, T. (1994). A critique of pure vision. In Koch, C. and Davis, J. L., editors, *Large scale neuronal theories of the brain*, pages 23–60. MIT Press, Cambridge, MA.

Clark, A. (1997a). *Being There: putting brain, body and world together again*. Oxford University Press, Oxford.

Clark, A. (1997b). The dynamical challenge. *Cognitive Science*, 21(4):461–481.

Clark, A. (2003). Forces, fields, and the role of knowledge in action. *Adaptive Behavior*, 11:270–272.

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, 10(8):370–374.

Clark, A. and Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior*, 7(1):5–16.

Clark, A. and Toribio, J. (1994). Doing without representing? *Synthese*, 101:401–431.

Cliff, D. and Noble, J. (1997). Knowledge-based vision and simple visual machines. *Philosophical Transactions to the Royal Society of London B*, 352:1165–1175.

Clowes, R. and Morse, A. F. (2005). Scaffolding cognition with words. In Berthouze, L., Kaplan, F., Kozima, H., Yano, H., Konczak, J., Metta, G., Nadel, J., Sandini, G., Stojanov, G., and Balkenius, C., editors, *Proceedings Fifth International Workshop on Epigenetic Robotics:Modeling Cognitive Development in Robotic Systems*, pages 101–105, Lund. Lund University Cognitive Studies.

Cummins, R. (1989). *Meaning and Mental Representation*. MIT Press, Cambridge, MA.

Dennett, D. C. (1978). *Brainstorms*. MIT Press, Cambridge, MA.

Dennett, D. C. (1991). *Consciousness Explained*. Little Brown & Co., New York, NY.

Dennett, D. C. (1993). Learning and labeling. *Mind and language*, 8(4):540–547.

Dennett, D. C. (1994). Artificial life as philosophy. *Artificial Life*, 1(1):291–292.

Dietrich, E. and Markman, A. (2003). Discrete thoughts: Why cognition must use discrete representations. *Mind and Language*, 18:95–119.

Dretske, F. (1988). *Explaining behavior*. MIT Press, Cambridge, MA.

Dreyfus, H. L. (2002). Intelligence without representation merleau-ponty's critique of mental representation the relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences*, 1(4):367–383.

Edelman, S. (2003). But will it scale up? not without representations. *Adaptive Behavior*, 11:273–275.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Findlay, J. M. and Gilchrist, I. D. (2003). *Active Vision. The Psychologyof Looking and Seeing*. Oxford University Press, Oxford.

Floreano, D., Kato, T., Marocco, D., and Sauser, E. (2004). Coevolution of active vision and feature selection. *Biological Cybernetics*, 90(3):218–228.

Fodor, J. A. (1981). *Representations: Philosophical Essays on the Foundations of Cognitive Science*. MIT Press, Cambridge, MA.

Gallagher, S. (2008). Are minimal representations still representations? *International Journal of Philosophical Studies*, 16(3):351–69.

Garzon, F. C. (2008). Towards a general theory of antirepresentationalism. *Br J Philos Sci*, 59(3):259–292.

Gentner, D. (2003). Why we are so smart. In Gentner, D. and Goldin-Meadow, S., editors, *Language in mind*, pages 195–235. MIT Press, Cambridge, MA.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin, Boston.

Gigliotta, O. and Nolfi, S. (2008). On the coupling between agent internal and agent/environmental dynamics: Development of spatial representations in evolving autonomous robots. *Adaptive Behavior*, 16:148–165.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, 20(01):1–55.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral And Brain Sciences*, 27(3):377–396.

Harnad, S. (2005). To cognize is to categorize: Cognition is categorization. In Cohen, H. and Lefebvre, C., editors, *Handbook of Categorization in Cognitive Science*, pages 20–45. Elsevier, Oxford.

Harvey, I. (1996). Untimed and misrepresented: connectionism and the computer metaphor. *AISB Quarterly*, 96:20–27.

Harvey, I., Husbands, P., and Cliff, D. (1994). Seeing the light: artificial evolution, real vision. In *From animals to animats 3: Proceedings of the third international conference on Simulation of adaptive behavior*, pages 392–401, Cambridge, MA. MIT Press.

Haselager, P., De Groot, A., and van Rappard, H. (2003). Representationalism vs. anti-representationalism: a debate for the sake of appearance. 16(1):5–24.

Haugeland, J. (1991). Semantic engines: An introduction to mind design. In Haugeland, J., editor, *Mind Design: Philosophy, Psychology, Artificial Intelligence*, pages 1–34. MIT Press, Cambridge, MA.

Hesslow, G. (2002). Conscious thought as simulation of behavior and perception. *Trends in Cognitive Sciences*, 6(6):242–247.

Jeannerod, M. (2006). *Motor Cognition: What Actions Tell to the Self*. Oxford University Press, Oxford.

Jeannerod, M., Arbib, M. A., Rizzolatti, G., and Sakata, H. (1995). Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neuroscience*, 18(7):314–320.

Keijzer, F. A. (2001). *Representation and Behavior*. MIT Press, Cambridge, MA.

Kelso, J. (1995). *Dynamic Patterns*. MIT Press, Cambridge, MA.

Lupyan, G. (2005). Carving nature at its joints and carving joints into nature: How labels augment category representations. In Cangelosi, A. and Bugmann, G.a nd Borisyuk, R., editors, *Modelling Language, Cognition and Action: Proceedings of the 9th Neural Computation and Psychology Workshop*, pages 87–96, Singapore. World Scientific.

Markman, A. and Dietrich, E. (2000a). In defense of representation. *Cognitive Psychology*, 40:138–171.

Markman, A. B. and Dietrich, E. (2000b). Extending the classical view of representation. *Trends in Cognitive Sciences*, 4(12):470–475.

Meyer, J.-A. and Wilson, S. W., editors (1990). *From animals to animats: Proceedings of the first international conference on simulation of adaptive behavior*, Cambridge, MA, USA. MIT Press.

Millikan, R. G. (1984). *Language, thought and other biological categories*. MIT Press, Cambridge, MA.

Millikan, R. G. (1996). Pushmi-pullyu representations. In *Philosophical Perspectives*, volume 9, pages 185–200. Ridgeview Publishing.

Mirolli, M., Ferrauto, T., and Nolfi, S. (2010). Categorisation through evidence accumulation in an active vision system. *Connection Science*, 22(4):331–354.

Mirolli, M. and Parisi, D. (2005a). How can we explain the emergence of a language which benefits the hearer but not the speaker? *Connection Science*, 17(3-4):325–341.

Mirolli, M. and Parisi, D. (2005b). Language as an aid to categorization: A neural network model of early language acquisition. In Cangelosi, A., Bugmann, G., and Borisyuk, R., editors, *Modelling language, cognition and action: Proceedings of the 9th Neural Computation and Psychology Workshop*, pages 97–106, Singapore. World Scientific.

Mirolli, M. and Parisi, D. (2006). Talking to oneself as a selective pressure for the emergence of language. In Cangelosi, A., Smith, A., and Smith, K., editors, *The Evolution of Language: Proceedings of the 6th International Conference on the Evolution of Language*, pages 214–221. World Scientific Publishing.

Mirolli, M. and Parisi, D. (2009). Language as a cognitive tool. *Minds and Machines*, 19(4):517–528.

Mirolli, M. and Parisi, D. (2010). Towards a vygotskyan cognitive robotics: The role of language as a cognitive tool. *New Ideas in Psychology*.

Murata, A., Fadiga, L., Fogassi, L., Gallese, V., Raos, V., and Rizzolatti, G. (1997). Object representation in the ventral premotor cortex (area f5) of the monkey. *Journal of Neurophysiology*, 78(4):2226–2230.

Noë, A. (2004). *Action in Perception*. MIT Press, Cambridge, MA.

Nolfi, S. (1997). Evolving non-trivial behavior on autonomous robots: Adaptation is more powerful than decomposition and integration. In Gomi, T., editor, *Evolutionary Robotics*, pages 21–48. AAI Books, Ontario (Canada).

Nolfi, S. (1998). Evolutionary robotics: Exploiting the full power of self-organization. *Connection Science*, 10(3-4):167–183.

Nolfi, S. (2002). Power and limits of reactive agents. *Neurocomputing*, 49:119–145.

Nolfi, S. and Marocco, D. (2001). Evolving robots able to integrate sensory-motor information over time. *Theory in Biosciences*, 120(3):287–310.

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

Pezzulo, G. (2008). Coordinating with the future: the anticipatory nature of representation. *Minds and Machines*, 18:179–225.

Pfeifer, R. and Bongard, J. C. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press.

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

Port, R. F. and van Gelder, T., editors (1995). *Mind as Motion*. MIT Press, Cambridge, MA.

Prinz, J. and Barsalou, L. (2000). Steering a course for embodied representation. In Dietrich, E. and Markman, A., editors, *Cognitive dynamics: Conceptual change in humans and machines*, pages 51–77. MIT Press, Cambridge, MA.

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge University Press, Cambridge.

Rizzolatti, G., Camarda, R., Fogassi, M., Gentilucci, M., Luppino, G., and Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey: Ii. area f5 and the control of distal movements. *Experimental Brain Research*, 71:491–507.

Rizzolatti, G. and Sinigaglia, C. (2008). *Mirrors in the Brain. How our Minds Share Actions and Emotions*. Oxford University Press, Oxford.

Rumelhart, D., Hintont, G., and Williams, R. (1986a). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Rumelhart, D., McClelland, J., and the PDP Research Group (1986b). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1 & 2. MIT Press, Cambridge, MA.

Scheier, C., Pfeifer, R., and Kunyioshi, Y. (1998). Embedded neural networks: exploiting constraints. *Neural Network*, 11(7-8):1551–1569.

Schyns, P. G. (1991). A modular neural network model of concept acquisition. *Cognitive Science*, 15(4):461–508.

Singer, W. (1999). Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24(1):49–65.

Smith, B. C. (1996). *On the Origin of Objects*. MIT Press, Cambridge, MA.

Spelke, Elizabeth, S. (2003). What makes us smart? core knowledge and natural language. In Gentner, D. and Goldin-Meadow, S., editors, *Language in mind*, pages 277–311. MIT Press, Cambridge, MA.

Spencer, J. P. and Schoner, G. (2003). Bridging the representational gap in the dynamic systems approach to development. *Developmental Science*, 6(4):392–412.

Steels, L. (2003a). Intelligence with representation. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 361(1811):2381–2395.

Steels, L. (2003b). Language-reentrance and the 'inner voice'. *Journal of Consciousness Studies*, 10(4-5):173–185.

Steels, L. and Brooks, R., editors (1994). *The artificial life route to artificial intelligence: Building Situated Embodied Agents*. Lawrence Erlbaum Ass., New Haven.

Sugita, Y. and Tani, J. (2005). Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adaptive Behavior*, 13(1):33–52.

Thelen, E. and Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, MA.

Tomasello, M. (2003). The key is social cognition. In Gentner, D. and Goldin-Meadow, S., editors, *Language in mind*, pages 47–57. MIT Press, Cambridge, MA.

Tucker, M. and Ellis, R. (2001). The potentiation of grasp types during visual object categorization. *Visual Cognition*, 8:769–800.

van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, 92(7):345–381.

van Gelder, T. J. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21:615–665.

van Gelder, T. J. (1999). Dynamic approaches to cognition. In Wilson, R. and Keil, F., editors, *The MIT Encyclopedia of Cognitive Sciences*, pages 243–245. MIT Press, Cambridge MA.

VanRullen, R., Guyonneau, R., and Thorpe, S. J. (2005). Spike times make sense. *Trends in Neurosciences*, 28(1):1–4.

Varela, F., Thompson, E., and Rosch, E. (1991). *The Embodied Mind*. MIT Press, Cambridge, MA.

Vygotsky, L. S. (1978). *Mind in society*. Harvard University Press, Cambridge, MA.

Wheeler, M. (1994). From activation to activity. *AISB Quarterly*, 87:36–42.

Wheeler, M. (2005). Friends reunited? evolutionary robotics and representational explanation. *Artif. Life*, 11(1-2):215–232.

Yarbus, A. L. (1967). *Eye Movements and Vision*. Plenum Press, New York.