Are There Representations in Embodied Evolved Agents? Taking Measures

Hezi Avraham[1], Gal Chechik[2], and Eytan Ruppin[1,3]

[1] School of Computer Sciences,
Tel-Aviv University, Tel-Aviv 69978, Israel,
{hezuz,ruppin}@post.tau.ac.il
[2] The Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem, 91904, Israel,
ggal@cs.huji.ac.il
[3] School of Medicine,
Tel-Aviv University, Tel-Aviv 69978, Israel

Abstract. The question of conceptual representation has received considerable attention in philosophy, neuroscience and embodied evolved agents. Numerous theories on the interpretation of the term 'representation' exist, and many arguments have been made for and against the existence of representations in animate and animat agents. Our work studies this question in evolved artificial embodied agents in a quantitatively rigorous manner, for the first time. We develop two measures, based on information theory, to account for representations. These measures are studied by applying them to evolved agents performing a visual categorization, generalized XOR task. Our results show that having quantitative measures still leaves one with arbitrary "threshold values" decisions which permit wide freedom in determining the existence of representations. However, and more importantly, our results show that information-theoretic measures can still be used efficiently to identify discriminative neural patterns and internal structures that characterize a representation, if the latter is formed.

1 Introduction

Internal representations are thought to play a central role in our understanding of cognitive behavior and information processing of intelligent agents. Yet, the idea that representations actually exist in the brains of animate or animat agents has been seriously challenged by many. For example, Brooks [1] has proposed a bottom-up approach to building agents that are able to react in real time in their environment, while having no central model representing the world. Cliff and Noble [2] point to the lack of evidence of representations in artificial evolved systems solving simple visual tasks, where the term 'representation' denotes a sub-network's pattern of activity which marks an external object or event to the rest of the network. Rather much like Pfeifer and Scheier [3], they suggested that the workings of agents could be best understood by studying the dynamics of

W. Banzhaf et al. (Eds.): ECAL 2003, LNAI 2801, pp. 743–752, 2003.

sensory-motor coordination between the agent and the environment. The usefulness of the notion of representations has also been a focus of much debate in cognitive neuroscience, e.g., [4,5,6].

One classical way to search for representations in cognitive neuroscience research is to look for an emergence of a "correlational structure" in the patterns of neural activity studied [4,6]. Intuitively, this is tantamount to the assertion that representation of an object corresponds to neural activity that marks its existence, but is indifferent to the different instances of it. Another interesting approach, termed in information theoretic notions, has been suggested by Usher [7], where Shannon mutual information is used to characterize the information that a concept/representation carries about external items. These investigations, however, have studied representations in standard connectionist networks, or have remained on a conceptual and philosophical level. **Given this long-time and fundamental controversy in cognitive neuroscience, the present work studies, in a quantitative manner, the question of representation in embodied evolved agents**. The agents studied evolve to solve a generalized XOR task by successfully categorizing visual stimuli. This task was chosen because it is known as a non-trivial evolutionary benchmark challenge [8], and because it is characterized by vanishing correlations in the sensory input readings, on the background of which the formation of a structure in neural activity is necessary and may be readily assessed. Moreover, by this we avoid wrong conclusions using neural networks, which can exploit first-order correlations between input and output and by that ignore higher-order configurational information [5].

Our goals in this work are twofold: One is to gauge whether representations are indeed formed in our agents. The other more important one is to capitalize on the simplicity and transparency of evolved agent models and develop rigorous and quantitative measures of "representation". The latter should serve as working rulers in future studies of this fundamental issue. The paper is organized as follows. We begin by describing the model and experimental protocol by which we evolve the agents. In Sect. 3 we present two quantitative approaches for defining and measuring representations, and describe the results obtained when these methods are applied to the evolved agents. Finally, in Sect. 4, we discuss the implications of our findings for measuring and understanding representation in embodied agents.

2 Methods

The model environment (*world*) is a $14.0 \times 14.0 \times 2$ three dimensional semi continuous simulated arena, where the arena's length and width are continuous while the height is discrete. The world contains two types of *resources*, 10 *food* items and 10 *poison* items. The resources are randomly scattered in a 10.0×10.0 central *resources zone*. Each resource type has a shape comprised of two circular pellets (each pellet radius can be either 0.02 or 0.08), placed one on top of the other with a common center point (Fig. 1a,c). The agent behavioral task is eating

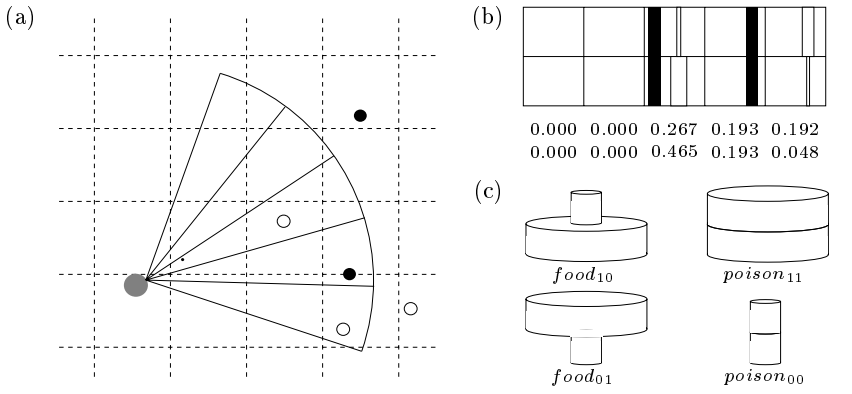

Fig. 1. The model outline. (a) A top view of the simulated world. An agent (*gray circle*) looking at two food items (*white circles*) and at two poison items (*black circles*) through a 5 sub-slices scan sensor. (b) The projected sensor retina (*box*) along with the photoreceptors values (*numbers*). (c) Resources shapes in the generalized XOR task

as much food as possible while avoiding poison. The agent is initially placed at a random location in the arena and eats a resource by simply colliding into it.

The agent consists of three main systems: a *scan sensor*, constantly giving readings of the environment; a synchronous continuous feed-forward neural network, processing the sensor input and producing two output signals; and a motor system, comprising of two independent wheels on either side of the agent's cylindrical body (0.15 radius), each driven by one output motor neuron. The sensor is mounted in front of the agent's body and processes a 3D distal resource stimuli to form a 2D retinal image. The sensor looks ahead for a distance of 3, acceptance angle of 90 ° and a height of 2. The sensor retina is realized as an array of 10 photoreceptors (Fig. 1b), each activated by the existence of a resource in the corresponding sensor's *sub-slice*. A photoreceptor value is set to the amount of total projections from its corresponding sub-slice, yielding a real number in the range of [0,1] (Fig. 1a,b).

A synchronous continuous feed-forward neural network realizes the agent's neurocontroller. We use several network architectures with 1-2 hidden layers. The network acquires its inputs from 10 dedicated sensory neurons, each transmitting the value of one photoreceptor. The output layer is composed of two output neurons, which control the agent's wheels. All network neurons set their activities according to: $V_i(t) = g(h_i(t))$, where $V_i(t)$ and $h_i(t)$ denote neuron i's activity and field (input vector sum) at network update t correspondingly, and $g(x)$ is the sigmoid function $g(x) = 1/(1 + e^{-\beta(x-\theta_i)})$, where $\beta = 4$ is the squashing factor and θ_i is a bias term, set in the genome for each neuron.

The agent motion is determined by its speed and orientation which are set by the wheels velocities and their difference, correspondingly. The agent speed is between -1 and 1 *world units* per time step, and its turn angle is in the interval of [-36 °,36 °].

An agent's lifetime lasts 50 sensory-motor cycles, followed by a normalized fitness score calculated as the number of food items it has consumed minus the number of poison items it has eaten. An evolution run lasts 10,000 generations. In each generation a population of 100 agents is evaluated, then the parents of the next generation are chosen with probability proportional to the agents fitness. At the end of an evolution session the precise fitness of an agent is assessed over 1,000 epochs. The agent's genetic encoding is a string of real valued numbers describing the neurocontroller synaptic weights and the neurons' bias terms θ_i. The initial genes values are randomly chosen in the range of -1 and 1, and have no limits during the evolution run. The genetic operators employed are uniform crossover with probability of 0.35 and mutation with probability of 0.02 and range of [-0.6,0.6].

The experiment we conducted (*generalized XOR* task) is designed to challenge the agents with a 'visual' generalized XOR problem. 5 food items have a common shape of large pellet (0.08) placed underneath a small one (0.02) ($food_{10}$). The remaining 5 food items have the opposite arrangement ($food_{01}$). Similarly, there are two types of poison: 5 poison items have large lower and upper pellets ($poison_{11}$), while the rest of the poison items have small pellets ($poison_{00}$) (Fig. 1c). Hence, the agent has to solve the generalized XOR problem in order to distinguish successfully food from poison. We evolve two kinds of neurocontrollers, one with a single hidden layer with 5 neurons (XOR-*5-2*), and another with two hidden layers, the first with 6 neurons and the second with 4 neurons (XOR-*6-4-2*).

3 Results

3.1 Performance

Direct evolution failed to come up with a good solution for the generalized XOR problem (fitness of 0.2421 and 0.2563 for XOR-*5-2* and XOR-*6-4-2*, correspondingly[1]), therefore we used incremental evolution [9]. This is executed by having large pellets with radiuses of 0.1 (instead of 0.08) and small pellets with a significantly reduced size in the initial stage of the evolution, and then gradually modifying the pellets sizes to their original values as the incremental evolution successfully progresses. Indeed, this technique succeeds and proper agents are evolved (fitness of 0.3275 and 0.3770 for XOR-*5-2* and XOR-*6-4-2*, correspondingly). Further analysis focuses on these two agents, which exhibit a well adjusted behavior, circling the arena in order to keep within the resource zone, avoiding poison, turning to food on either side and then moving with full speed ahead towards it.

[1] The maximal fitness feasible can be roughly estimated by viewing an agent evolved in a *poison-less* world with a single food type. Here, the best evolved individual attains fitness of 0.4272, having a neurocontroller of one hidden layer with 4 neurons.

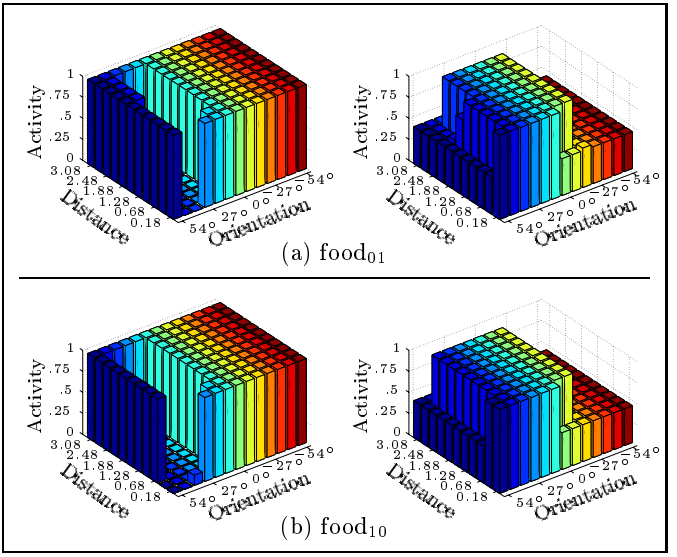

Fig. 2. Left and right output motor neurons' receptive fields in response to food stimuli, for agent XOR-*6-4-2*'s. (a) Left and right responses to $food_{01}$. (b) Left and right responses to $food_{10}$. Each bar marks the neuron's activation level in response to a food stimuli in a particular distance and orientation

3.2 Behavioral Analysis

A first clue to representations in embodied agents can be obtained from a behavioral analysis. If distinct behavioral responses are found to different types of food (or poison), the search for food representation, that is indifferent of food type, turns irrelevant since representation is not used to constitute an akin behavior. Conversely, if similar behavioral responses to both types of a resource are observed, it supports the possibility that a joint internal representation is formed.

In order to compare the agent behavior to the different resources types, we measured the output neurons' receptive fields. This was done by recording their activation levels while introducing, one at a time, all possible resource instances in the agent's field of view (110 discrete locations for each resource type). As can be clearly seen (Fig. 2), there is significant similarity in agent XOR-*6-4-2*'s output neurons receptive fields, both with regard to food types and with regard to poison types (not shown). This affirms that the agent acts similarly in response to the different types of food and similarly for the different types of poison (an analogous conclusion is made regarding agent XOR-*5-2*).

3.3 Entropy in Embedded Agents

An ideal method for studying the relationship between objects and their neural representations, in face of a stochastic noisy environment, is information theory.

It provides a rigorous and quantitative approach to measure the high order correlations that are typical to the type of non linear processing performed by neurocontrollers. Usher [7] has suggested a quantification of representation that is based on the mutual information between a coherent state of the system and the input stimuli in the environment. Following this approach, we develop a novel information related measure of representations in agents. Let us denote $S \in \{s_{\mathrm{f}}, s_{\mathrm{p}}\}$ where $S = s_{\mathrm{f}}$ if the object stimuli is food and $S = s_{\mathrm{p}}$ if it is poison. Similarly, denote $F \in \{f_{01}, f_{10}\}$ and $P \in \{p_{00}, p_{11}\}$. Let $R = \{r_1, \ldots, r_n\}$ denote the set of n coherent states of the neural system examined. To identify these states in the experiments reported below, we applied a K-means algorithm to the neural activities recorded over 2000 agent's life trials (epochs), using a Euclidean norm. The r_i's are obtained as the centroids of neural activities clusters. Each neural activity vector is tagged by the most central object in the agent's view, e.g., as food_{10} or poison_{11}. Using this notation we calculate the conditional entropy of S given R [10]

$$H(S|R) = \sum_{i=1}^{n} p(r_i) H(S|R = r_i) = -\sum_{i=1}^{n} \sum_{j=\mathrm{f,p}} p(r_i) p(s_j|r_i) \log_2 p(s_j|r_i) \ . \quad (1)$$

This measure expresses the uncertainty of knowing whether food or poison was spotted given the activity pattern of the *sub-network* examined. It obtains values in the range of 0 to 1. If food and poison yield distinct activity patterns, $H(S|R)$ will obtain vanishing values. On the other hand, similar activity patterns for both food and poison would yield high $H(S|R)$ values close to 1.

Importantly, this measure alone cannot account for a representation since low $H(S|R)$ values may be obtained when activity patterns for the two food types (or poison types) are different. Since this case does not capture the notion of representation, $H(S|R)$ must be complemented with a measure of the difference in activity patterns within different types of food or poison. Therefore, food representation should be described as the difference between $H(S|R)$ and the entropy $H(F|R)$, computed likewise using food instances only. Since the representation of food (poison) may lay in the activity of a subset of the neurons examined, we define the *Euclidean Entropy Measures* (EEM) for the subset T,

$$EEM_{\mathrm{f}}(T) = H(F|R_T) - H(S|R_T) \ \text{ and } \ EEM_{\mathrm{p}}(T) = H(P|R_T) - H(S|R_T) \quad (2)$$

where R_T is a set of clusters obtained using K-means with Euclidean norm over the neuronal activities of all neurons in the subset T. This measure can take values in the range of $[-1, 1]$. It obtains a value of 1 when the responses for the different types of food (poison) are identical and the responses for food are completely different from those to poison, therefore denoting that a true representation of food (or poison) has been formed. A value of -1 is given when there are indistinguishable neural activities for food and for poison but distinct activities between both types of food (poison). In the case of random activity patterns, a value of zero is obtained.

Figure 3 shows the values of EEM calculated for all relevant sub-networks of the XOR-*6-4-2*'s hidden layers. It shows that EEM_{p} values are relatively constant

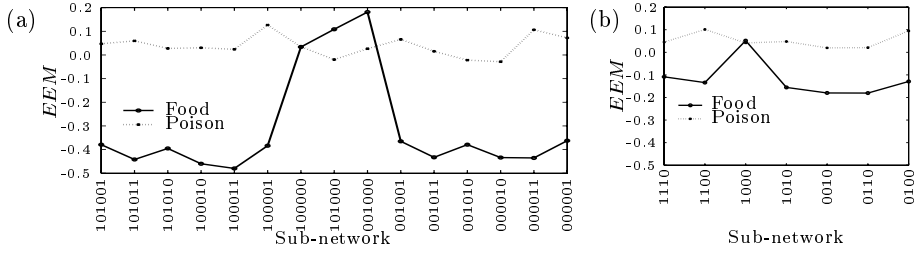

Fig. 3. Euclidean Entropy Measures for agent XOR-*6-4-2*. EEM_{f} (*solid*) and EEM_{p} (*dotted*) for (a) first hidden layer and (b) second hidden layer. A sub-network label (*x-axes*) denotes its configuration, where 1 means that the corresponding neuron exists in the sub-network. Since some neurons have a constant activity (i.e., neurons 2 and 4 in the first hidden layer and neuron 4 in the second hidden layer), only the relevant subsets for each layer are presented

across all subsets, yielding a distributed poison representation. On the other hand, EEM_{f} values are more variable and gain their maximum in few subsets, each having only few neurons, indicating a more localized food representation (qualitatively similar results were obtained for agent XOR-*5-2*).

In order to provide an index for each layer and each resource we further define

$$EEM^{*} = \max_{T} EEM(T) \ , \tag{3}$$

to measure the maximal EEM over all possible subsets. We calculated four EEM^{*} values for agent XOR-*6-4-2* (for food and for poison representations in each of its two hidden layers), which are generally low (0.181 for food and 0.127 for poison in the first layer, and 0.051 for food and 0.101 for poison in the second layer). This results from high $H(F|R)$ and $H(P|R)$ values but high $H(S|R)$ values (in the range [0.79,0.94]), indicating almost optimal food or poison unification but a poor food-poison separation.

The first hidden layer EEM^{*} values are higher than those of the second hidden layer, especially for food. One may speculate that the second hidden layer is strongly affected by the motoric constraints (e.g., same left turn for food on the left and for poison in the center), thus resource representation is more likely to form in the first hidden layer, yielding these higher EEM^{*} values.

3.4 Extracting Representations Using Information Bottleneck with Side Information

EEM assumes that the representation lies in a Euclidean metric on the single neuron firing patterns. **An alternative approach, is to cluster patterns of activities based on their functional relevance.** This approach was formally defined and analyzed in [11], using the Information Bottleneck method with Side Information ($IBSI$). $IBSI$ allows to search for structures in neural activities that are relevant to the discrimination between food and poison, but

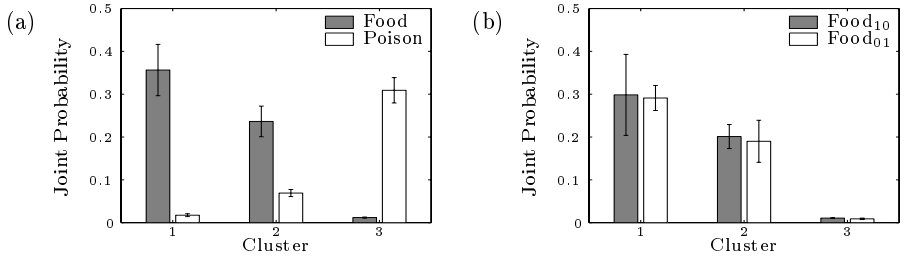

Fig. 4. *IBSI* joint probabilities. (a) $P(S, R)$ and (b) $P(F, R)$ means and SEM calculated on agent XOR-*6-4-2*'s first hidden layer. Clusters were obtained by applying the sequential-IBSI hard clustering algorithm with $\gamma = 1$ and *number of clusters*= 3. Similar results were obtained for wide range of γ values and number of clusters

not to the discrimination between *different types* of food or poison. IBSI can be thought of as a clustering analysis procedure that operates on the conditional distributions $P(S|R = r)$ and $P(F|R = r)$ rather than on the neural activities R. It therefore pre-assumes no particular metric between neural activity patterns at the clustering phase, and thus allows to test various kinds of metrics, rather than to use them as the basis for clustering.

Applied to the problem at hand, $IBSI$ searches for clusters that compress well a discrete set of neural activities X into a set of clusters R, while maintaining information about S and removing information about F (similarly about P),

$$IBSI_{\mathrm{f}} = \min_{p(r|x)} I(X;R) - \beta[I(R;S) - \gamma I(R;F)] \ , \tag{4}$$

where β and γ are tradeoff parameters functioning as Lagrange multipliers (see [11] for details). Inspection of the resulting set of clustered distributions $P(S, R)$ and $P(F, R)$ reveals a clear discrimination between food and poison (Fig. 4a) while hardly providing any knowledge about the different types of food (Fig. 4b).

It should be noted that both our EEM^* and $IBSI$ measures calculate very similar information theoretic measures[2]. However, EEM^* is obtained by maximizing over clusters that are formed using an Euclidean measure over neural activities, while $IBSI$ is obtained by directly maximizing the weighted difference of informations. To compare the two measures on the same scale, we computed the difference of entropies used for EEM (Eq. 2), over the clusters obtained from $IBSI$ (where T is the whole layer). $IBSI$ achieved higher values than those obtained with EEM (0.5898 for food and 0.5990 for poison in XOR-*6-4-2*'s first hidden layer, and 0.2246 for food and 0.2049 in the second hidden layer). One hypothesis regarding the large values observed in the first hidden layer is that it uses another (non-Euclidean) metric to solve the generalized XOR task while the second hidden layer metric is more similar to the Euclidean metric since

[2] In fact for $\gamma = 1$ and $\beta \to \infty$ these measures become equivalent up to a constant $H(F) - H(S)$.

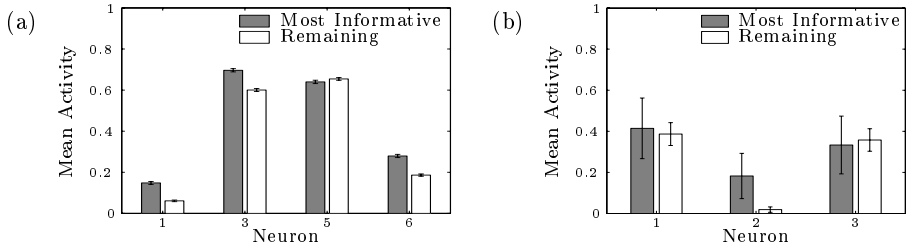

Fig. 5. *IBSI* food representations of agent XOR-*6-4-2*. Mean and *SEM* of the neural activities in the most informative cluster (*gray bars*) and in the remaining clusters (*white bars*) for (a) the first hidden layer and for (b) the second hidden layer. Only neurons with variable (and hence informative) activity are presented

is must refer to the motor actions to be executed. Another possibility is that the second hidden layer values bounds are smaller since the same motor actions should result from food and poison stimuli at different parts of the visual field.

After *IBSI* was used to extract functionally relevant clusters of neural activities, we turn to identify the nature of food and poison representations in the hidden layers neurons. To this end, we focus on the most informative cluster, defined as the cluster with majority of food stimulated responses, that maximizes the *single-symbol information* $D_{KL}[p(s|r)||p(s)] - D_{KL}[p(f|r)||p(f)]$ (and similarly for poison) where D_{KL} is the *Kullback-Liebler divergence* [10]. We first compare the mean neural activities in this cluster with the remaining activities. Figure 5 demonstrates this comparison for agent XOR-*6-4-2*, and reveals that **the mean activities of neurons 1, 3 and 6 in the first hidden layer and neuron 2 in the second hidden layer are significantly different in this cluster than in the remaining ones** (p <0.001 using t-test). This indicates that the activities of these isolated neurons is discriminative (qualitatively analogous results are obtained for agent XOR-*5-2*). Moreover, **several pairs of neurons were found to have cluster specific correlations** that were not observed in their baseline activities, which were found to be statistically significant using standard statistical test for correlation difference (for example, correlation coefficients of r_1=0.485 and r_2=-0.075 for neurons 2 and 3 in agent XOR-*6-4-2*'s second hidden layer, yielding p of 0.0156). This suggests that food representations lies not only in single neurons activities but also in function specific correlations.

4 Conclusions

This study raises two major questions: a) Are representations created in embodied evolved agents? and b) Can these representations be rigorously defined and measured? To answer these questions, we define two measures, *EEM* and *IBSI*-based, aimed to quantitatively and rigorously score inner representations. These two measures differ in their fundamental premises; the first assumes an

Euclidean metric between the neural activities, while the latter has no specific underlying metric assumption. The EEM measure produces low values for all agents and layers studied while $IBSI$ values are significantly higher, but firm conclusions about the existence or non-existence of representations remain subject to arbitrary "threshold value" decisions. However, we demonstrate that using these quantitative approaches leads to the identification of several important representational characteristics, including localized and distributed structures, discriminative neural activity patterns and cross-neuronal interactions. While much remains to be done in future studies, this work clearly shows that it's high time to corroborate the ongoing important conceptual debate about representations with a rigorous, quantitative investigation.

Acknowledgments. We acknowledge the valuable contributions made by Isaac Meilijson and Alon Keinan and the technical help provided by Oran Singer. This research has been supported by the Adams Super Center for Brain Studies in Tel Aviv University and by the Israel Science Foundation founded by the Israel Academy of Sciences and Humanities. G.C. is supported by the Israeli Ministry of Science.

References

1. Brooks, R.A.: Intelligence without representation. Artificial Intelligence **47** (1991) 139–159
2. Cliff, D., Noble, J.: Knowledge-based vision and simple visual machines. Philosophical Transactions of the Royal Society: Biological Sciences **352** (1997) 1165–1175
3. Pfeifer, R., Scheier, C.: Sensory-motor coordination: the metaphor and beyond. Robotics and Autonomous Systems, Special Issue on "Practice and Future of Autonomous Agents" **20** (1997) 157–178
4. Kosslyn, S.M., Chabris, C.F., Marsolek, C.J., Koenig, O.: Categorical versus coordinate spatial relations: computational analyses and computer simulations. Journal of Experimental Psychology: Human Perception and Performance **18** (1992) 562–577
5. Cook, N.D.: Correlations between input and output units in neural networks. Cognitive Science **19** (1995) 563–574
6. Kosslyn, S., Chabris, C., Baker, D.: Neural network models as evidence for different types of visual representations. Cognitive Science **19** (1995) 575–579
7. Usher, M.: A statistical referential theory of content: Using information theory to account for misrepresentation. Mind and Language **16** (2001) 311–334
8. Stanley, K.O., Miikkulainen, R.: Evolving neural networks through augmenting topologies. Evolutionary Computation **10** (2002) 99–127
9. Gomez, F., Miikkulainen, R.: Incremental evolution of complex general behavior. Adaptive Behavior **5** (1997) 317–342
10. Cover, T., Thomas, J.: The elements of information theory. Plenum Press, New York (1991)
11. Chechik, G., Tishby, N.: Extracting relevant structures with side information. In Becker, S., Thrun, S., Obermayer, K., eds.: Advances in Neural Information Processing Systems 15 (NIPS-2002), MIT press (2003)