Neuromodulation in a learning robot: Interactions between neural plasticity and behavior

Olaf Sporns, William H. Alexander
Department of Psychology, Indiana University
Bloomington, IN 47405

Abstract-**We investigate the effects of behavioral activity on the development of patterns of synaptic connections in a model of neuromodulation embedded in an autonomous robot. Robot behavior produces clustered spatial distributions of rewarding objects in environments initially containing a high density of uniformly distributed objects. The spatial clustering of the rewarding objects restricts robot movements to sectors of high reward density, alters patterns of connections signaling predictions about reward timing, and changes the temporal profile of interactions between the robot and the objects. These effects are outcomes of embodied behavioral activity and are not pre-programmed or externally controlled. We discuss our results in the context of the reciprocal coupling of neural dynamics, behavioral activity and stimulus distributions.**

INTRODUCTION

Neuromodulatory systems have powerful effects on learning and plasticity throughout the brain. They have been demonstrated to influence a wide range of neural function, including neuronal activity, synaptic strength, neuronal morphology and patterns of gene expression. Their important role in shaping behavior makes them potential key components of neural control architectures embedded in autonomous robots.

In a series of previous studies [1,2,3], we designed and implemented a neural network model of the mammalian mesencephalic dopamine system, which was designed to investigate the function of dopamine in the context of behavior. The model was based on the anatomy (neural connectivity) and physiology of dopamine neurons in the ventral tegmental area (VTA) and their associated cortical networks. We focused on the highly characteristic temporal patterns of activation observed in the VTA in response to unpredicted or predicted primary rewards [4,5]. Dopamine neurons discharge immediately after an unpredicted reward is first encountered. This phasic response is then transferred in the course of learning to other stimuli that reliably predict the occurrence of the reward. These observations have been modeled in the context of temporal difference learning [6,7] and have given rise to the hypothesis that midbrain dopamine neurons encode positive and negative prediction errors, i.e. change their firing rate in accordance with whether or not a future reward is predicted or whether a predicted award is omitted [5,6,7]. Our neural network model was able to show a broad spectrum of phenomena, including: a) dopamine responses to unpredicted rewards; b) transfer of the dopamine response to reward-predicting sensory stimuli (as a result of synaptic plasticity in afferents to the dopamine system itself); c) inhibition (attenuation) of the dopamine response at the time of omission of a previously predicted reward stimulus d) dopamine responses to stimuli that occur either early or late; and e) extinction.

Our model was tested in computer simulation [3] and in an autonomous robot [2,3]. When using discrete learning trials, during which stimuli and rewards occurred at specific time intervals, both computer simulation and experiments conducted with the robot yielded results that were consistent with the observed physiological characteristics of the mammalian dopamine system. However, little experimental data is available on the performance and temporal characteristics of the dopamine system in the context of free, unconstrained and autonomous behavior in any mammalian species. This paper represents the continuation of an earlier study [3] aiming at characterizing the interplay between "internal" neural dynamics and plasticity and "external" factors such as motor variables and environmental composition. Our initial results suggested that behavioral activity has several unforeseen consequences on the learning process itself and impacts on the specific pattern or synaptic connectivity that develops within the neural model of the robot. Here, we further explore this aspect of our model and present additional evidence for the close and reciprocal interactions between neural plasticity and behavioral activity. These interactions form an important rationale for conceptualizing and studying development and learning as an embodied process.

METHODS

The robot, environment and neural model used and implemented in the present paper are described in detail in a previous publication [3, see also ref. 2]. Here, we provide only a brief overview of the model's main features.

General. All data presented in this paper have been collected using a mobile robotic platform (Khepera robot), interfaced with a neural simulation run on a workstation, and showing autonomous behavior in an environmental enclosure containing colored (red) objects. All neural and synaptic states are recorded for each experiment and analyzed and displayed off-line. In addition, an overhead camera records views of the environment and of the robot's actions at a rate of approximately 1 frame/second.

Robot. The robot (named "Monad") is capable of navigating through the environment, avoiding obstacles (walls) guided by IR sensors, and of physically contacting objects using a moveable arm and gripper. Main sensors interfaced with the neural model are vision (through a color CCD camera mounted on the robotic platform) and "taste" (conductivity of objects). Low conductivity activated appetitive taste receptors, signaling the occurrence of a reward.

Environment. The enclosure was about 1 meter on each side and contained red (appetitive) objects. These objects were placed in the environment by the experimenter at the beginning of each run. The objects were light and could easily be moved and pushed around by Monad in the course of an experiment. We conducted experiments with low object density (4 objects total) and high object density (12 objects total).

Behavior. Visual detection of objects (described in detail in [2]) resulted in approach behavior under visual guidance. Once an object was in close proximity, the arm was lowered and the gripper closed, while Monad turned slightly to align the gripper surfaces with the object (see movies at http://php.indiana.edu/~osporns/embodied.htm). Sensing of the "taste" of the object was followed by its subsequent release and the continuation of environmental exploration. The encounter of reward in conjunction with the gripping of red objects strengthened connections between appropriate visual and motor units, rendering vision capable of triggering the behavioral sequence leading to physical contact with ("consumption of") red objects. Since only rewarding objects were used in the present set of experiments, there is little or no overt change in behavior (i.e. the unconditioned response is very similar to the conditioned response after learning). Note that other experiments [2] using aversive objects yielded different behavioral patterns (approach and avoidance) in the course of learning.

Neural Model. All neural units were implemented using a continuous firing rate model with a single saturating non-linearity, according to

$$s_i(t+1) = \phi[A(t) + \Omega s_i(t)] \qquad (1)$$

where $s_i(t)$ is the activity of unit i at time t, $A(t)$ is the total synaptic input to unit i at time t, Ω is the unit's temporal persistence ($0<\Omega<1$), and $\phi[.]$ is the saturating nonlinearity (both hyperbolic tangent and sigmoidal functions are used). $A(t)$ was calculated as the linear sum of all inhibitory and excitatory inputs, i.e. $\Sigma c_{ij} s_j(t)$.

The neural model of the mammalian midbrain dopamine system was aiming to incorporate the functional and anatomical characteristics of the real neural structures. In the neural model, several types of cells comprised an area analogous to the ventral tegmental area (VTA), whose outputs modulated plasticity in several different locations. Input to the modeled VTA was provided by a temporal representation of the visual stimulus (representing an analogue to prefrontal cortex, PFC). Color selective units (C_{red}) relayed sensory inputs to a network transforming this input into a continuous temporal representation. As a result of excitatory and inhibitory interactions within this network (essentially forming a delay chain), stimulus-specific units D_{red} became active after a specific amount of time (between 1 and 12 iterations) had elapsed since the onset of their preferred stimulus. These units had fairly broad temporal tuning with significant mutual overlap in terms of their "temporal receptive field". D_{red} units projected to the VTA (for more detail on the structure of the VTA and its connectivity, see [2,3]), making a series of 12 modifiable connections (strengths are shown in plots in Fig. 3 and 4). The overall level of neuromodulator released by the VTA reward system was taken to be a "value signal" or reward signal V_R. V_R was a signed scalar variable, which was positive if the level of neuromodulator increased above a baseline and negative if it decreased below a baseline. The neural model used this value signal in value-dependent learning to adjust all connection strengths. Through value-dependent learning, the value signal V_R influenced synaptic modification in three sets of connections. Of particular importance in this paper are the 12 connections linking the temporal delay units D_{red} to the VTA. Through value-dependent modification of these connections, value (reward) was able to modify the response characteristics of parts of the neuromodulatory system (the VTA) itself. Connections that were subject to value-dependent learning were updated according to a ternary learning rule:

$$c_{ij}(t+1) = (1-\varepsilon)c_{ij}(t) + \eta s_j(t)F(s_i(t))V_R \qquad (2)$$

where c_{ij} = connection weight from unit j to unit i, ε = incremental decay rate of connection weight per iteration, η = learning rate, F(.) = nonlinear function applied to postsynaptic activity, V_R = value signal. F(.) determined if a connection weight increased or decreased, depending upon the level of postsynaptic activity. F(.) was a continuous saturating function ($-1<F(.)<1$) with a temporal profile that reflected the dependence of synaptic modifications on postsynaptic activity.

RESULTS

We focus on a detailed analysis of the development of synaptic and environmental variables in the course of unconstrained autonomous behavior. Experimental runs are carried out under conditions of high object density (12 objects) with an initially uniform spatial distribution of

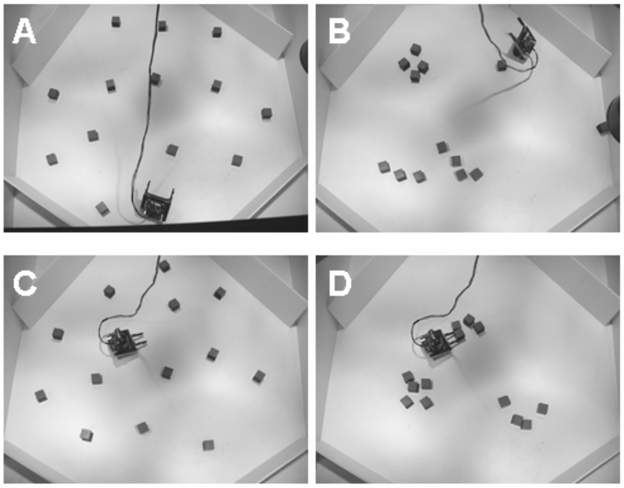

Fig. 1. Distribution of objects before (A, C) and after (B, D) an experimental run with high object density. A, B and C, D show views from two separate experiments.

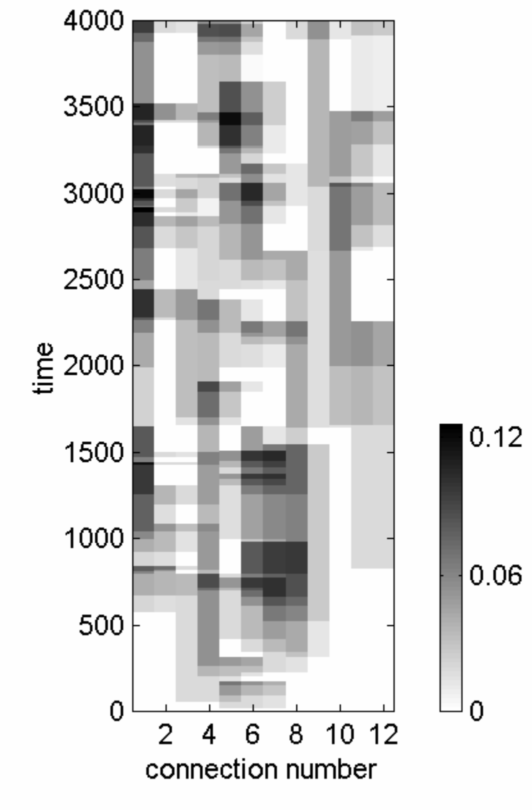

Fig. 2. Development of synaptic weights between D_{red} and the VTA model, during a single experiment (see Fig. 1 A,B). Connection number refers to the 12 connections mediating temporal delays of between 1 and 12 iterations (see text for detail). Time (4,000 iterations total) starts with t=0 at the bottom of the plot. The gray scale at the right indicates synaptic weight.

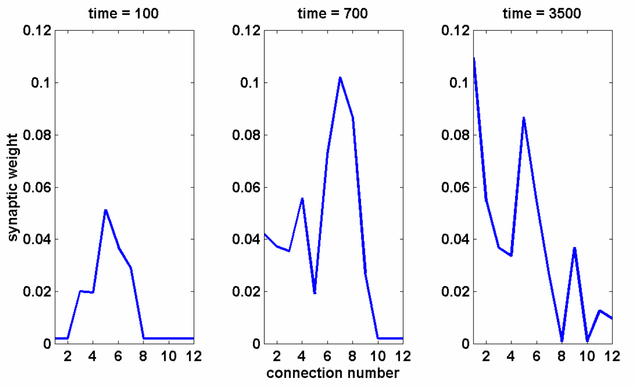

Fig. 3. Weight profiles obtained after 100 (left), 700 (middle) and 3500 iterations (right), from the same experiment shown in Fig. 2.

objects in the environment. All objects are rewarding (appetitive). Experiments lasted for 3,000 to 4,000 iterations (about 10 to 15 minutes of real time).

Fig. 1 shows the distribution of the objects at the beginning and at the end of two representative experiments. Note that the objects are displaced from their original positions and are pushed together into clusters as a result of the behavioral and motor activity of the robot. Thus, the initially uniform spatial distribution of reward in the environment is changed into one that consists of "clusters" or "islands" of high density of reward, surrounded by areas devoid of rewarding objects. This tendency of mobile robots to generate uneven object distributions in their environments as a result of behavior was noted earlier in other neuro-robotic experiments (see Discussion). No such clustering occurs if the initial density of objects is low (4 objects, see ref. 3).

Fig. 2 shows the profile of synaptic weights of connections linking visual delay units (D_{red}) and the dopamine system. In timed trials these connections display specific patterns encoding information about the relative timing of the onset of reward-predicting stimuli and the reward itself [2,3]. However, in the course of autonomous behavior the timing between the initial visual detection of a red object and the delivery of the reward (taste after gripping) is less precise and often inconsistent from one encounter of an object to the next. This degree of variability is due to two main factors: a) high variance in the distance between robot and object when the object is first encountered; b) and uneven distribution of objects due to due to their spatial rearrangement into clusters and islands. Consequently, the synaptic patterns in Fig. 2 do not reflect a single consistent temporal delay between object vision and reward. Rather, they incorporate a complex mix of delays, especially towards the end of the experiment.

Fig. 3 shows samples of the weight profiles at 100, 700 and 3,500 iterations during a representative experiment. Initially, a consistent temporal prediction emerges (Fig. 3,

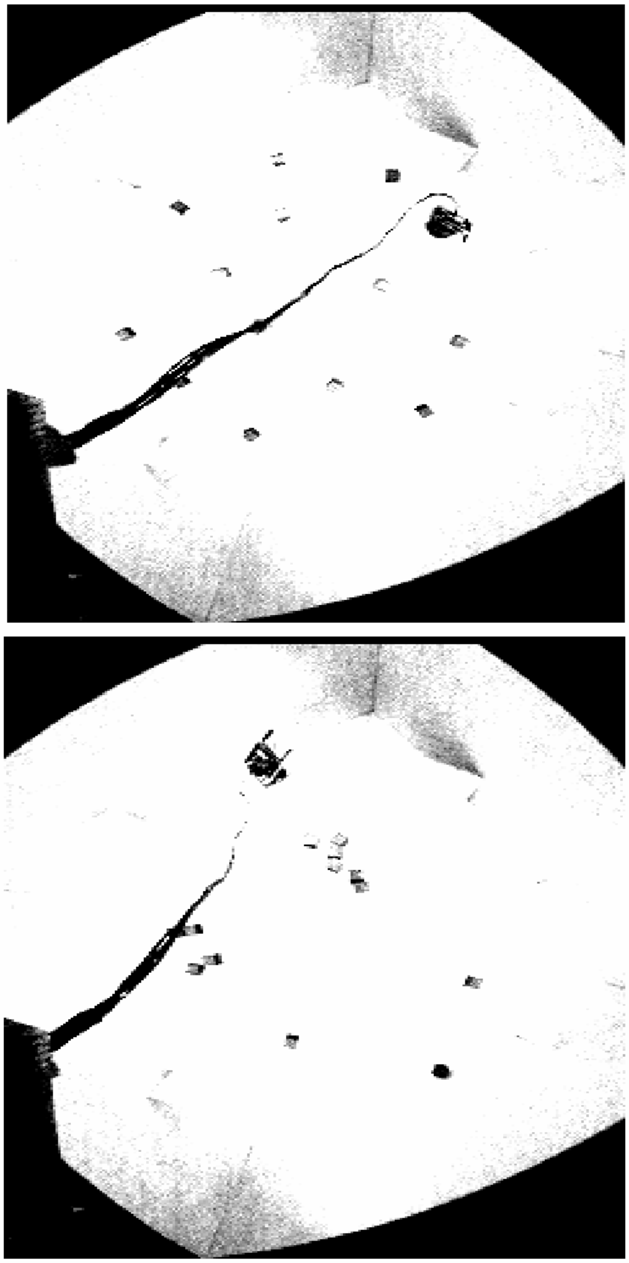

Fig. 4. Views of the environment from an overhead camera used to record behavior. Top: Beginning of an experiment involving 12 objects. Bottom: End of the same experiment. Note again how objects are redistributed and clustered together (see also Fig. 1)

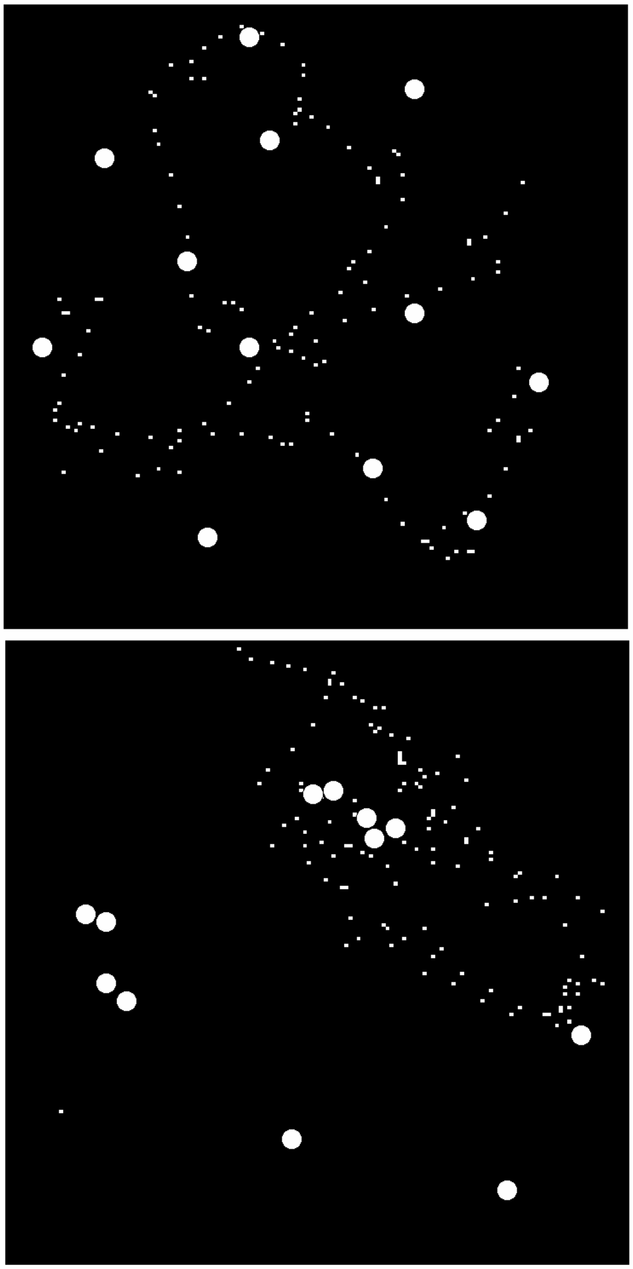

Fig. 5. Spatial locations occupied by Monad (small dots) for a period of approximately 700 iterations at the beginning (top) and end (bottom) of a representative experiment (same views as central portion of Fig. 4). Object locations are marked by white circles.

left) and is consolidated into a single dominant peak in the weight distribution (Fig. 3, left and middle). Towards the end of the experiment, however, this initial distribution has given way to a multi-peak spectrum of synaptic weights, seemingly reflecting the superposition of multiple temporal predictions (Fig. 3, right). Particularly prominent is an early peak (connection number 1) indicating that reward is often encountered immediately after visual contact is made.

To explain this pattern of synaptic weights, we plot the spatial distribution of objects and locations/trajectories of the robot during early and late phases of a representative experiment. Fig. 4 shows overhead camera views of Monad and the 12 objects within its environment. Fig. 4 (top) shows the initial distribution (uniform) and Fig. 4 (bottom) shows the distribution after approximately 3,500 iterations. As in Fig. 1, characteristic clustering of objects, due to the activity of Monad, is apparent. Fig. 5 displays the locations occupied by Monad during two periods of 600 iterations at the very beginning of the experiment (Fig. 5, top) and at the end of the experiment (Fig. 5, bottom;

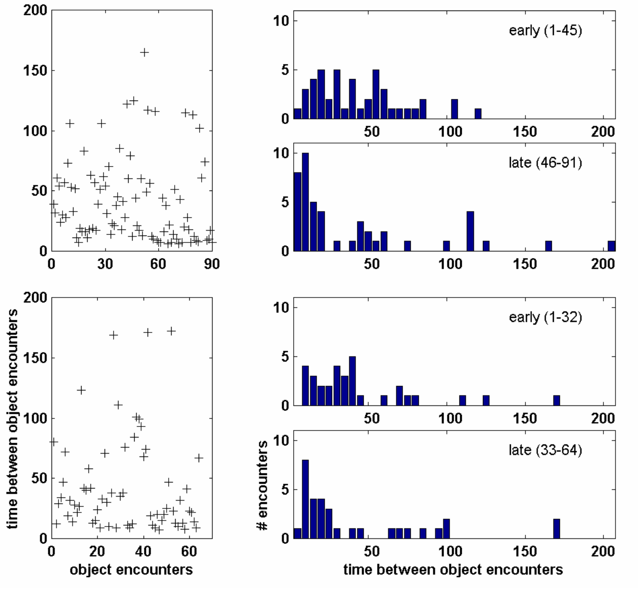

Fig. 6. Left: plots of time between object encounters over the course of two separate experiments (see Fig. 1). Right: Histograms of times between object encounters for early and late phases during both experiments. Data from experiment 1 is summarized in the top three plots (91 object encounters), data from experiment 2 is summarized in the bottom three plots (64 object encounters). The median time between successive object encounters for "early (1-45)" (summarizing the first 45 object encounters for experiment 1) is 39, the median for "late (45-91)" is 17.5. The corresponding medians for the lower two histograms are 34 and 21, respectively.

both views correspond to those shown in Fig. 4). At the beginning of the experiment, Monad traverses large regions of the environment, navigating from object to object and interacting with them individually. Objects are well separated, such that visual approach to most objects commences as the object enters the periphery of the robot's visual field. This results in fairly predictable temporal delays between first visual contact and subsequent reward. This pattern of behavior allows the early emergence of a consistent prediction in the afferent synaptic weights to the dopamine system (cf. Fig. 3, left and middle). Later in the experiment, objects have been redistributed, due to physical interactions between objects and robot. Monad tends to be strongly "attracted" to object clusters, resulting in several sequences of behavioral interactions with rewarding objects that occur in rapid temporal succession. As shown in Fig. 5 (bottom), Monads trajectory is largely restricted to the vicinity of one of these object clusters. A statistical analysis of the timings of successive reward occurrences reveals that significant changes in reward timing have indeed taken place, when comparing early and late phases of the experiments (Fig. 6). In early phases, reward encounters are, on average, temporally more spread out, while in later phases, many rewards are presented in rapid succession, with longer time intervals between these sequences reflecting navigation between object clusters.

Discussion

Our experiments document a progressive alteration of an environmental variable (the spatial distribution of reward throughout the environment) due to the behavioral activity of the robot. This alteration, in turn, has consequences on synaptic patterns encoding predictions about the occurrence of future rewards.

It is especially noteworthy that the differences between early and late phases in experiments with high object densities are neither the result of purposeful rearrangements of the environment by either robot or experimenter, nor are they due to the adjustment of "internal" variables over time such as learning rates, cell response functions, or motor variables. Instead they are the outcome of the coupling between brain, body and environment. This coupling is strongly reciprocal. Behavior affects the statistics of reward timings which drive synaptic plasticity through activation of a neuromodulatory system. In turn, synaptic changes alter the coupling between visual and motor units which affects behavior.

Other robot models have demonstrated the role of behavior in shaping development and learning. The emergence of specific receptive field properties of modeled neurons in the inferotemporal cortex was shown to depend on the continuity and smoothness of robot movements [8]. Perceptual categorization was facilitated by robot motor activity [9,10] in different contexts. It was also highly sensitive to the statistical distribution of stimuli and the developmental history of stimulus encounters [8,11]. Numerous lines of experimental evidence from human and animal studies suggest a critical role of bodily movement in cognitive and neural development.

The experiments discussed in this paper may shed light on the activity and functional role of neuromodulatory systems (in particular, dopamine) in the course of "natural", self-guided behavior. The "attractive force" exerted by clusters of rewarding objects, resulting in restricted trajectories of robot movement and navigation as well as repeated "rapid-fire" sequences of reward encounters are especially intriguing. Disruptions of the neurobiological bases of reward processing are thought to form a major cause for lasting behavioral changes and, eventually, chronic disease (addiction) in humans. Our results suggest the hypothesis that a pattern of persistent reward-seeking behavior may in part be generated as a result of a progressive reshaping of the environment coupled with long-lasting synaptic changes in specific neural structures. Future experiments will investigate this hypothesis in detail.

The effects of behavior and motor activity on shaping the statistics of sensory inputs provide a fundamental theoretical argument for the importance of *embodied* cognition [12,13] The observations reported in this paper depended critically on behavior carried out in a real environment. They could not have been predicted from the consideration of timed trials alone or from computer simulations that do not incorporate the coupling between brain and behavior. Future work will focus on the development of appropriate behavioral and neural measures that, in combination, characterize mental and cognitive development [14]. Further analyses are needed to elucidate the nature of brain/behavior interactions and to identify and develop appropriate statistical methods that capture their complexity.

Acknowledgment

This work was supported by NIH/NIDA grant 1R21DA15647-01.

References

[1] Sporns, O., Almassy, N., and Edelman, G.M. (2000) Plasticity in value systems and its role in adaptive behavior. *Adaptive Behavior*, 8, 129-148.

[2] Sporns, O., and Alexander, W.H. (2002) Neuromodulation and plasticity in an autonomous robot. *Neural Networks,* 15, 761-774.

[3] Alexander, W.H., and Sporns, O. (2003) An embodied model of learning, plasticity and reward. Adaptive Behavior, 10, (in press).

[4] Schultz, W. (1998) Predictive reward signal of dopamine neurons. Journal of Neurophysiology, 80, 1-27.

[5] Schultz, W., Dayan, P. and Montague, P.R. (1997) A neural substrate of prediction and reward. Science, 275, 1593-1599.

[6] Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996) A framework for mesencephalic dopamine systems based on predictive hebbian learning. Journal of Neuroscience, 16, 1936-1947.

[7] Suri, R.E. (2002) TD models of reward predictive responses in dopamine neurons. Neural Networks, 15, 523-533.

[8] Almassy, N., Edelman, G.M., and Sporns, O. (1998) Behavioral constraints in the development of neuronal properties: A cortical model embedded in a real world device. Cerebral Cortex, 8, 346-361.

[9] Scheier, C. and Lambrinos, D. (1996) Categorization in a real-world agent using haptic exploration and active perception. In From animals to animats: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior, pp. 65-75, Eds. Maes, P., Mataric, M., Meyer, J.-A., Pollack, J., Wilson, S.W., MIT press, Cambridge, MA.

[10] Pfeifer, R., and Scheier, C. (1998) Representation in natural and artificial agents: An embodied cognitive science perspective. Zeitschr. Naturforschung, 53c, 480-503.

[11] Krichmar, J.L.; Snook, J.A.; Edelman, G.M.; Sporns, O. (2000) Experience-dependent perceptual categorization in a behaving real-world device. In: Animals to Animats 6: Proceedings of the Sixth International Conference on the Simulation of Adaptive Behavior, Meyer, J.A.; Berthoz, A.; Floreano, D.; Roitblat, H.; Wilson, S.W., (Editors), MIT Press: Cambridge, MA. p. 41-50.

[12] Sporns, O. (2002) Embodied cognition. In: *Handbook of Brain Theory and Neural Networks, 2nd Edition.*, Arbib, M. (ed.), Cambridge, MA: MIT Press, pp. 395-398.

[13] Pfeifer, R., and Scheier, C. (1999) Understanding Intelligence. MIT Press, Cambridge, MA.

[14] Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., Thelen, E. (2001) Autonomous mental development by robots and animals. *Science*, 291, 599-600.