Contents lists available at SciVerse ScienceDirect

Journal of Physiology - Paris

journal homepage: www.elsevier.com/locate/jphysparis

# Reinforcement active learning in the vibrissae system: Optimal object localization

Goren Gordon [a,*], Nimrod Dorfman [b], Ehud Ahissar [a]

[a] Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel
[b] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel

ARTICLE INFO

*Article history:*
Available online 9 July 2012

*Keywords:*
Intrinsic reward
Whiskers
Curiosity loop
Perception
Palpation

ABSTRACT

Rats move their whiskers to acquire information about their environment. It has been observed that they palpate novel objects and objects they are required to localize in space. We analyze whisker-based object localization using two complementary paradigms, namely, active learning and intrinsic-reward reinforcement learning. Active learning algorithms select the next training samples according to the hypothesized solution in order to better discriminate between correct and incorrect labels. Intrinsic-reward reinforcement learning uses prediction errors as the reward to an actor-critic design, such that behavior converges to the one that optimizes the learning process. We show that in the context of object localization, the two paradigms result in palpation whisking as their respective optimal solution. These results suggest that rats may employ principles of active learning and/or intrinsic reward in tactile exploration and can guide future research to seek the underlying neuronal mechanisms that implement them. Furthermore, these paradigms are easily transferable to biomimetic whisker-based artificial sensors and can improve the active exploration of their environment.

## 1. Introduction

Rats are curious animals that use their vibrissae (whiskers) to explore their environment. Several stereotypical behaviors have been observed, such as periodic whisking (Gao et al., 2001; Berg and Kleinfeld, 2003) and touch-induced palpation (Grant et al., 2009). Recently, whisking behavior has been implemented in robotic whiskers in order to discriminate textures and ascertain three-dimensional shapes (Solomon and Hartmann, 2006; Evans et al., 2010; Sullivan et al., 2012). Palpation of novel objects, which is the focus of the current work, is observed when rats encounter such an object and can be characterized as a high-frequency small-amplitude whisker motion, always remaining in the vicinity of the object. It has received very little attention from the analytical and robotics-implementation fields (Gordon and Ahissar, 2011; Gordon and Ahissar, 2012).

Here we show that two seemingly unrelated paradigms, namely, active learning (Kolodziejski et al., 2009; Bhatnagar et al., 2007; Govindhasamy et al., 2005) and intrinsic-reward reinforcement learning (Barto et al., 2004; Weng, 2004; Oudeyer et al., 2007; Schmidhuber, 2010), predict that touch-induced palpation is the optimal behavior for whisker-based object localization. We then show that in the context of object localization, the two paradigms are tightly related and suggest neuronal mechanisms that may implement each.

Rats' vibrissae system serves as a unique model for neuroscience research due to its relative simplicity. Although its dynamics becomes more complex as investigations progress (Knutsen and Ahissar, 2009; Simony et al., 2010), it can be approximated as a one-dimensional process, controlling a single positional variable, the whisker's azimuth angle using a single motor variable, whisker velocity. Then whisker-based object localization can be defined as learning the forward model (Jordan, 1992; Shadmehr and Krakauer, 2008) of touch, i.e. the ability to predict at what angle and velocity a touch signal, due to contact between the whisker and object, will occur. The question we address is "how should a single-whisker rat move its whisker in order to optimally localize an object?" In other words, what is the rat's policy that optimizes learning of the forward model of touch, where optimization is performed with respect to the learned function (see below).

This scenario can be formulated using the active learning jargon in the following way (Adejumo and Engelbrecht, 1999; Dasgupta and Hsu, 2008). The rat *samples* the sensory-motor space (angle and velocity) and wishes to correctly *label* each point as touch or no-touch. We show that object localization is equivalent to learning a two-dimensional linear separator (albeit in bounded space due to angle and velocity limitations). Hence, the goal is to find the sampling policy that minimizes the error between the predicted linear separator and the correct one.

In reinforcement learning (RL) notations (Kolodziejski et al., 2009; Bhatnagar et al., 2007; Govindhasamy et al., 2005), the *states*

* Corresponding author. Tel.: +972 525374429; fax: +972 775374424.
*E-mail addresses:* goren@gorengordon.com (G. Gordon), Ehud.Ahissar@weizmann.ac.il (E. Ahissar).

0928-4257/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jphysparis.2012.06.004

are the angle of the whisker and the touch information, and the *action* is whisker velocity. Hence in an actor-critic setup (Bhatnagar et al., 2007), the critic learns the values of each angle/touch point, whereas the actor adjusts the probabilities of choosing a specific whisker velocity given an angle/touch state. In conventional RL, the reward is given by an *extrinsic* function that is adjusted to the desired goal, e.g. maximal reward for arriving at a specific location. However, in the current implementation of RL, the object localization component, i.e. a learner that learns the forward model of touch, provides *intrinsic* reward (Barto et al., 2004; Weng, 2004; Oudeyer et al., 2007; Schmidhuber, 2010), here taken to be the prediction error. Thus, the goal is to find the actor that optimizes learning object localization, i.e. minimizes the generalization error of the forward model of touch. In other words, the intrinsic-reward RL converges to a behavior that results in fast increase in accurate prediction of touch events.

The paper provides a unifying formalism for both approaches, with respect to object localization via whiskers. This allows the direct comparison of the two approaches, which exhibit remarkably similar results, namely, palpation behavior. It further enables a formulation of the connection between the two paradigms, also explored here. A biologically-plausible neuronal network that implements the proposed models is also presented and discussed. Finally, the synergistic analysis presented here can facilitate the application of either techniques in robotic whisker-base sensors (Solomon and Hartmann, 2006; Evans et al., 2010; Sullivan et al., 2012).

## 2. Materials and methods

### 2.1. Whisker model

We use a simplistic model, in which the rat can *control the velocity* of the whisker (Simony et al., 2010). Furthermore, the whisker itself is rigid, i.e. it cannot bend and hence its azimuth angle cannot pass the “object’s angle”. Thus, the whisker *angle is bounded* by the object position and depends on the initial whisker angle, i.e. if it is initially more retracted (smaller angle) or more protracted (larger angle) than the object (angular) position. For simplicity, we assume the whisker to be always more retracted than the object; hence the object is touched only upon protraction of the whisker. This assumption is validated by numerous videos of freely moving rats, in which they encounter novel objects upon protraction in the vast majority of cases. We also assume that the velocity is bounded, due to physical constraints.

### 2.2. Learning a linear separator in sensory-motor space

We formulate the whisker-based object localization setup mathematically: $\theta \in \left[\theta^{\min}, b\right]$ is the angle of the whisker, $\theta^{\min}$ is the fully retracted angle, and $b$ is the (angular) position of the object, with $b \in \left[\theta^{\min}, \theta^{\max}\right]$, $\theta^{\max}$ being the fully protracted angle. This means that the object can appear anywhere inside the whisker field. We assume that the whisker is always more retracted than the angular position of the object, and hence bounded by it. $a \in \left[a^{\min}, a^{\max}\right]$ is the bounded velocity of the whisker.

The dynamics of the system are given by

$$\theta'_{t+1} = \theta_t + a_t \tag{1}$$

$$\theta_{t+1} = \max \theta_{\min}, \min(b, \theta'_{t+1})) \tag{2}$$

where $\theta'_{t+1}$ is the attempted angle and Eq. (1) guarantees that the angle stays within the bounds. The velocity $a_t$ is the action that should be optimized (see below). The touch signal is then given by

$$B_{t+1} = \begin{cases} 1 & \theta_t \leqslant b \text{ and } \theta'_{t+1} > b \\ -1 & \text{otherwise} \end{cases} \tag{3}$$

This means that if the whisker tried to move from one side of the object to the other side of the object, there is a touch signal of 1, otherwise $B = -1$. One can then define a linear separator of touch, $u = \{u_\theta, u_a, u_b\}$, such that

$$u_\theta \theta_t + u_a a - u_b = u^T x_t = 0 \tag{4}$$

where $x_t = \{\theta_t, a_t, -1\}$ is a point in 2-dimensional $(x^1 = \theta, x^2 = a)$ space, where $x^3 = -1$ is a constant added to accommodate for the linear separator’s threshold $u_b$. The linear separator, $u$, delineates the boundary between the labeled touch and no-touch regions in the two-dimensional $(\theta, a)$ space.

The setup can then be re-formulated as follows: (i) The agent’s policy determines, based on past knowledge, the action $a_t$; (ii) the dynamics are determined via Eqs. (1) and (3); (iii) the agent receives $\{\theta_{t+1}, B_{t+1}\}$; (iv) based on the action, angle and touch signal, the agent updates its approximation of the linear separator. (v) $t \to t+1$, return to (i). The goal is then restated as: find a policy such that the touch-signal linear separator, $u$, is learned optimally.

### 2.3. Perceptron-based active learning

The setup described in the previous section can be modeled by a perceptron, which is a mathematical construct that receives many inputs and has a single output. The perceptron output is the result of applying a (usually) non-linear or threshold function on the weighted sum of its inputs. In the object localization scenario, the perceptron inputs and output are the two-dimensional point ($\theta$, a) and touch signal, respectively.

The problem is also related to selective sampling, a branch of active learning (Settles, 2009), in which one can select whether to label the sample or not. Since the labeling is usually costly, the aim is to select which samples to label. We briefly describe a perceptron-based active learning algorithm taken from Dasgupta et al. (2009), which actively selects which samples to label and exhibits an exponential speedup compared to random selections.

Let $x$ be a point on the $N$-dimensional unit hypersphere, $\sum_{i=1}^{N} x_i^2 = 1$. Let $u$ be a vector on the same sphere, such that $y = \text{sign}(u^T x)$ is the label of each point $x$ on the sphere. In each time-step, $t$, there is a hypothesis vector, $v_t$. The goal of active learning is to find $u$, i.e. change the hypothesis such that $v_t \to u$.

In selective sampling, one is presented with random samples from the unit sphere, $x_t$. The algorithm presented in Dasgupta et al. (2009) only labels samples obeying:

$$|v_t^T x_t| < q_t \tag{5}$$

where $x_t$ is the sample at time $t$, $v_t$ is the current hypothesis/classifier and $q_t$ is an adaptive threshold that decreases as learning progresses. It was shown that the update rule of the hypothesis, $v_t$, given by

$$v_{t+1} = v_t - 2(v_t^T x_t) x_t \tag{6}$$

results in a number of required labels that is exponentially smaller for a given error, compared to random labeling. The crux of the algorithm in Dasgupta et al. (2009) is the adaptive threshold $q_t$, which adapts according to the following rule: if predictions were correct on $R$ consecutive labeled examples, then set $q_{t+1} = q_t/2$, else $q_{t+1} = q_t$. This means that the adaptive threshold decreases as the error in the prediction decreases.

### 2.4. Reinforcement active learning

Reinforcement learning (RL) deals with the question of finding an actor that maximizes (future) accumulated rewards. In our setup, which we call autonomous reinforcement active learning (ReAL), the goal is to optimize on-line supervised learning, in the sense that generalization error of the learner is minimized (Gordon and Ahissar, 2011, 2012). The heuristics behind ReAL is "you learn more by making (corrected) mistakes". To this end we define the reward to be the prediction error (Schmidhuber, 1990), i.e. the difference between the expected label and the correct label. This prediction error is used to adjust the learner parameters, but here it also serves as the intrisic reward signal (Barto et al., 2004; Weng, 2004; Oudeyer et al., 2007; Schmidhuber, 2010).

In the RL implementation, we set the states to be whisker angle and touch label, $s_t = \{\theta_t, B_t\}$, Eq. (3); and the actions are set to be whisker velocity, $a_t$. The object localization learner is the forward model of touch, i.e. predicting contact $B_{t+1}$, based on the current angle and velocity, $s_t, a_t$. It is denoted by $L(B_{t+1}|s_t, a_t)$ and is simulated in the ReAL model as a feed-forward multi-layer neural network with backpropagation learning algorithm. We implement the incremental Natural Actor Critic (iNAC) algorithm (Bhatnagar et al., 2007), and add an intrinsic reward that is the square of the forward model of touch prediction error:

$$r_{t+1} = [L(B_{t+1}|\theta_t, a_t) - B_{t+1}]^2 \tag{7}$$

We briefly summarize the iNAC algorithm below (Bhatnagar et al., 2007; Gordon and Ahissar, 2011, 2012). The rat selects an action, $a_t$, at each time $t$ using a randomized stationary policy, designated as the actor: $\pi(a|s) = \Pr(a_t = a|s_t = s; \lambda_t)$, where $\lambda_t$ are the actor parameters to be tuned. The Natural Actor-Critic algorithm uses the compatible functions, defined as: $\psi(s_t, a_t) = \nabla_\lambda \pi(a_t|s_t)$.

The critic, $\widehat{V}^\pi(s_t; v_t)$ attempts to learn the value function, i.e. the value of each state, by tuning the parameters $v_t$ using the function $\phi(s_t) = \nabla_v \widehat{V}^\pi(s_t; v_t)$. Moreover, the reinforcement learning algorithm uses the temporal difference (TD) learning, here taken to be $\delta_t = r_t - \widehat{J}_{t+1} + \widehat{V}^\pi(s_t; v_t) - \widehat{V}^\pi(s_{t+1}; v_t)$, where $\widehat{J}_t$ is the estimated average reward, which is also updated. The update rules are summarized below:

$$\widehat{J}_{t+1} = (1 - \xi_t)\widehat{J}_t + \xi_t r_t \tag{8}$$

$$v_{t+1} = v_t + \alpha_t \delta_t \phi(s_t) \tag{9}$$

$$w_{t+1} = \left[I - \alpha_t \psi(s_t, a_t)\psi(s_t, a_t)^T\right] w_t + \alpha_t \delta_t \psi(s_t, a_t) \tag{10}$$

$$\lambda_{t+1} = \lambda_t + \beta_t w_{t+1} \tag{11}$$

where $\xi_t$ is the average reward update rate, $w_t$ are the advantage parameters, $\alpha_t, \beta_t$ are the learning rates of the critic and actor, respectively, and their step-size schedule satisfy the condition that the critic converges faster than the actor (Bhatnagar et al., 2007).

In the ReAL algorithm, we have introduced a delicate interplay between three approximators, namely the actor, critic and learner. The actor, through the selection of the appropriate action and the state-change induced by the system dynamics, determines which new example is presented to the learner. This, in turn, produces the prediction error which not only modifies the learner weights, but also determines the reward, which the critic now assimilates into its value and advantage approximators. The critic completes the ReAL loop by determining the TD error that updates both the critic and the actor.

## 3. Results

### 3.1. Object localization via active learning

We first describe the modifications we implemented on the perceptron-based active learning algorithm described above (taken from Dasgupta et al. (2009)), in order to accommodate the whisker-based object localization setup. These are: (i) action-based active learning, also known as membership queries (Settles, 2009); (ii) the whisker space is not on the unit sphere, but is bounded both in the whisker angle and the whisker velocity Knutsen et al., 2008 and; (iii) only one dimension (velocity) is under complete control, while the other (angle) is only partially controlled through the velocity and the whisker dynamics, Eq. (1). We then present numerical simulations of the modified algorithm, exhibiting palpation behavior.

### 3.1.1. Action-based active learning

One defines action-based active learning as follows: all presented samples are labeled, with their respective costs, but one has some control over the next sample. Hence, the aim is to select such actions so as to sample next at the correct position. The analytical derivation is presented below.

In action-based active learning, one chooses an action that influences the next sample, such that it will fulfill Eq. (5), i.e. choose an action that generates

$$x_{t+1} = n_t + d_t \tag{12}$$

where $n_t$ is uniformly sampled from the space $\{n_t \in \Re^N | v_t^T n_t = 0, ||n_t||_2 = 1\}$ and $d_t$ is uniformly sampled from the space $\{d_t \in \Re^N | \, ||d_t||_2 \leqslant q_t\}$. Here $n_t$ denotes a random vector sampled from the hyperplane orthogonal to the current hypothesis, $v_t$, while $d_t$ denotes the distance from the hypothesis.

Eq. (12) means that one must find a random vector in the hyperplane orthogonal to the current hypothesis, with some added noise proportional to the adaptive threshold, $q_t$. This implies that instead of a random presentation of samples followed by labeling according to the selection condition, one performs an action such that the sample generated meets the selection condition.

### 3.1.2. Bounded space

The action-based adaptation was done straightforwardly in the unit sphere scenario. However, in the unit sphere, all samples lie on the sphere, by default. In the whisker-based object localization setup, both the whisker angle and velocity are bounded and do not obey the unit sphere constraint. Rather, they lie in a bounded region in the 2D plane. Hence, a new sample generated by Eq. (12), $x_{t+1} = \{\theta_{t+1}, a_{t+1}\}$, may fall outside the angle/velocity boundaries and must be modified to comply with those boundaries. This presents several new challenges.

The first is that the linear separator hypothesis does not necessarily pass through the origin, requiring learning an additional parameter, namely, the separator threshold, $u_b$. This is easily achieved by adding another auxiliary dimension to the linear separators, $v$ and $u$, as described in Eq. (4). Second, there is a possibility of getting stuck on the boundaries, due to an updated hypothesis, $v_{t+1}$ that lies entirely outside the bounded space. This is solved by taking the suggested next sample to be near the suggested separator, yet always within the bounded domain. Furthermore, only approximated separators that lie inside the space are allowed. Finally, if the separator is still stuck on the boundaries, one must restart the algorithm. While this increases the learning time, practically it happens rarely and does not reduce the algorithm's exponential speedup.

### 3.1.3. Velocity dependent control

In the angle/velocity scenario, only one dimension is under direct control (velocity) while the other (angle) is determined by the velocity and the dynamics. This actually simplifies the algorithm since the random vector $n_t$, drawn from the hyperplane orthogonal to the linear separator hypothesis, has only one free parameter instead of two. However, new pitfalls arise, namely, the required velocity $a_{t+1}$ can either lie outside the bounds or be zero. The former is overcome by restricting the velocity such that the next time-step whisker angle (according to Eq. (1)) lies within the angle bounds. The latter is overcome by prohibiting velocities below a certain threshold.

### *3.1.4. Numerical results*

In the object localization context, action-based active learning means that if the appropriate whisker motion is performed, the object, i.e. linear separator in the angle/velocity space, will be learned exponentially faster than random motion. The main feature of the algorithm in Dasgupta et al. (2009), i.e. the adaptive threshold, then amounts to palpation of the object, meaning motion that becomes closer and closer to the object boundary.

Fig. 1 shows an implementation of the algorithm, where the object is located at $b = 0$. Fig. 1a shows the object localization mapping, intensity coding the touch (white) and no touch (black) areas. The trajectory is marked in phase space by gray crosses, whose size is correlated to the time step (larger crosses, larger $t$). The trajectory starts randomly for a hundred time steps, Fig. 1c until it first reaches the object and then it "palpates" it by going back and forth between touching and not touching it. The learned separator converges from a random initial state (gray dashed line) to the correct separator (gray solid line), Fig. 1a, where Fig. 1b shows the exponential decrease in the separator error, computed as the distance between the true separator and approximated one, distance $= 1 - v \cdot u/|v||u|$. Fig. 1d shows the dynamical nature of the adaptive threshold.

### *3.2. Intrinsic-reward reinforcement learning of object localization*

In the incremental natural actor critic (iNAC) implementation of whisker object localization, we use the same touch signal as in Eq. (3), which may originate from a touch sensor (Gordon and Ahissar, 2011, 2012). The touch signal is binary, i.e. either there is touch with an object or there is not, where the strength of the touch that depends on its radial position along the whisker (Birdwell et al., 2007) and on the force of the whisker muscles (Simony et al., 2010), is neglected for simplicity. Furthermore, we have used a continuous state ($\theta$) and continuous action ($a$) RL algorithm, with an actor that depends on the touch information (and not the current angle) and on the previous action, $\pi(a_t|B_t, a_{t-1})$ (for full details, see Gordon and Ahissar (2012)).

Fig. 2a shows the object localization mapping for an object in the middle of the whisker field, similar to Fig. 1a. As can be seen it is confined to a small region around the object location and has a step-like shape, where far from the object there is no touch information and moving towards the object results in touch information, since the whisker motion is blocked by the object. In this example, the whisker always started from full retraction and the object was encountered during protraction. Hence, the generalization error of the mapping is computed only on one side that is determined by the initial state and the position of the object.

The actor used in this scenario was a non-Markov actor, in which the action depends on the current touch information and the previous action, Fig. 2b and c. Examining the trajectory of the learned actor, Fig. 2(e:upper, black), reveals that this actor *actively learns* the linear separator, similar to Fig. 1. This palpation whisking, i.e. alternating between touching and not touching the object, drastically increases the accumulated rewards (prediction errors) Fig. 2e:lower. The learning curve in Fig. 2d shows an initial worsening and then a drastic improvement in the generalization error. The former is due to the fact that the actor has learned to protract until an object is reached, avoiding the initial random exploration observed in Fig. 1c, which results in delayed learning of the large no-touch area. Following object-touch, which occurs after a small number of time steps, the palpation behavior learns the linear separator of the touch and no-touch areas and thus drastically reduces the generalization error. The random actor, on the other hand initially learns the large no-touch area, but due to a small number of touch events, fails to converge on the right linear separator that defines the object location.

Fig. 2b shows the action probabilities when there was no contact with the object and Fig. 2c when there was contact. This actor demonstrates a negative feed-back behavior, where it protracts

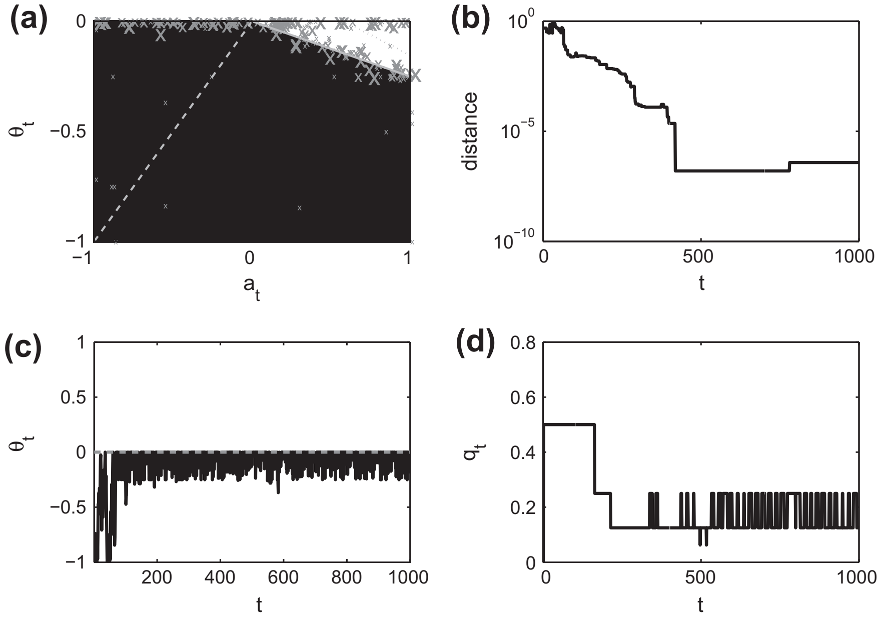

**Fig. 1.** Example of object localization as a 2D velocity dependent control of active learning. (a) Intensity coded map of true linear separator (black = $-1$, white = 1). Gray lines show approximated separator for initial (dashed), after 100 steps (dotted) and final after 1000 steps (solid). Gray crosses mark querying coordinates, where larger font-sizes indicates more advance time steps. (b) Distance measure between true and approximated separator. (c) Trajectory of whisker as a function of time, showing a change from semi-random wide-angle whisking to palaption whisking around the object. (d) Adaptive threshold as a function of time.

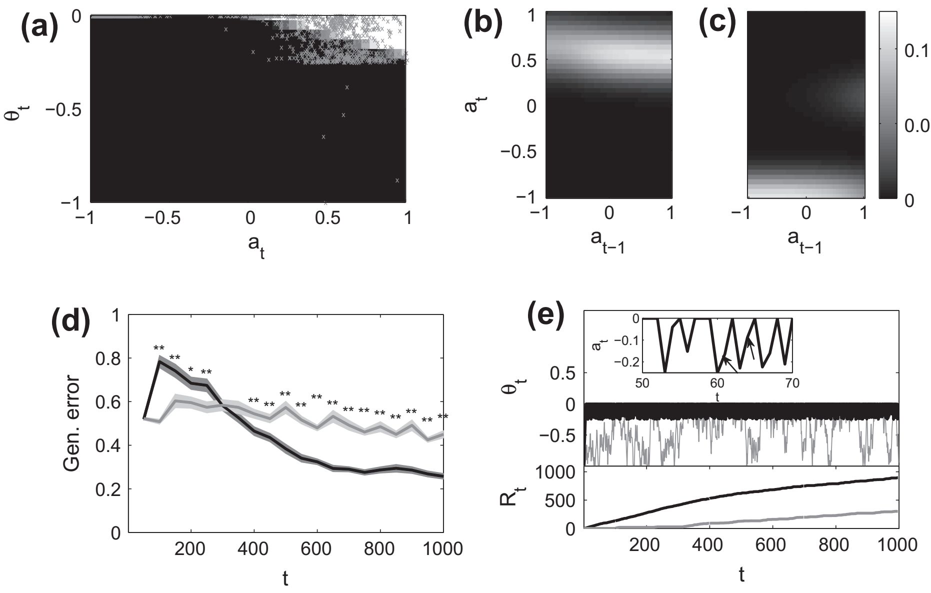

**Fig. 2.** ReAL of object localization. (a) Learned object localization mapping, B = 1 white, B = −1 black, gray crosses represent state-action trajectory. (b,c) Learned actor. (b) Action probabilities if no touch occurred. (c) Action probabilities if touch occurred. (d) Generalization error averaged over 100 actors (shaded areas represent standard error) for actors after 2000 ReAL episodes (black) and random actor (gray). (e) Upper panel: a typical trajectory of the learned (black) and random (gray) actors. Inset: part of the trajectories, arrows indicate non-Markov behavior. Lower panel: $R_t = \sum_{\tau=1}^{t} r_\tau$ is the accumulated reward of the same trajectory. Adapted with permission from Gordon and Ahissar (2012).

when not touching an object and retracts when it touches. However, the fact that the protraction is smaller than the retraction allows sampling the two sides of the linear separator, namely, after a large retraction follows two smaller protraction, the first without touch and the second with, indicated by the arrows in Fig. 2e:upper, inset. Furthermore, the probability of slightly protracting after a large protraction that induced touch, a non-Markovian feature, allows sampling of the most non-linear feature, namely, the intersection point of the linear separator with the object position at $\theta = b, a = 0$. Together, these unique features enable the active learning of the object localization linear separator.

### 3.3. Active learning and reinforcement learning with prediction error reward

We wish to connect the active learning setup as described above, to reinforcement learning in which the reward is proportional to the prediction error. Active learning consists of two steps, namely the "active" step in which the next sample is actively selected and the "learning" step in which the approximated separator is updated according to the sampled data. We show below that given the update rule, Eq. (6), expressed as an update of the approximated separator by a prediction or learning error $\epsilon_t$, the next sample that maximizes this prediction error obeys the active learning rule of sampling near the separator.

We wish to emphasize the maximization–minimization aspect of the connection between active and reinforcement learning: the update stage is designed to decrease the prediction error, whereas the active stage is shown below to select a sample that maximizes the prediction error. The rationale behind this architecture is that by actively going to places of maximal prediction error, their correction in the update stage will have the greatest effect. We begin from the update rule, given in (6), which is implemented only for a misclassification, i.e. when there is a prediction error given by:

$$\epsilon(x_t) = (v_t^T x_t)(u^T x_t) < 0 \tag{13}$$

where the $\widehat{B} = v_t^T x_t$ term is the (misclassified) prediction and the $B = u^T x_t$ term is the correct labeling. The goal of active learning is to choose the next sample $x_t$ such that $u^T v_{t+1} > u^T v_t$, i.e. the overlap between the hypothesis and true separators increases:

$$u^T v_{t+1} - u^T v_t = -2(v_t^T x_t)(u^T x_t) = 2|\epsilon(x_t)| > 0 \tag{14}$$

Restating this in the RL notation, one can set the reward to be the absolute value of the prediction error, $r_t = |\epsilon(x_t)|$. The goal is then to find the next sample that maximizes the reward:

$$\frac{\partial r_t}{\partial x_t^i} = \frac{\partial|\epsilon(x_t)|}{\partial x_t^i} = -(v_t^T x_t)u^i - v_t^i(u^T x_t) = 0$$
$$\Rightarrow u^i \sum_j v_t^j x_t^j + v_t^i \sum_j u^j x_t^j = 0 \tag{15}$$

Let us define $\rho_t$, which measures the distance of the sample from the hypothesized separator, and $\Delta_t$, the (unknown) error vector as:

$$\rho_t = v_t^T x_t \tag{16}$$

$$\Delta_t = u - v_t \tag{17}$$

Plugging these definitions into (15) results in:

$$v_t^i \sum_j \left(v_t^j + \Delta_t^j\right) x_t^j + \left(v_t^i + \Delta_t^i\right) \sum_j v_t^j x_t^j = 0 \Rightarrow \rho_t$$
$$= \frac{-\sum_j \Delta_t^j x_t^j}{2 + \sum_j \Delta_t^j / \sum_j v_t^j} \tag{18}$$

Hence, for the optimal next position, Eq. (18) suggests that the next sample should be dependent on the current error in approximation. For the case of bounded space considered here, we have $|x| \leqslant 1, |v| \leqslant 1$ that result in

$$|\rho_t| \leqslant \frac{\left|\sum_j \Delta_t^j\right|}{2 + \left|\sum_j \Delta_t^j\right|} \tag{19}$$

Furthermore, this solution indeed maximizes the update: $\frac{\partial^2 r_t}{\left(\partial x_t^i\right)^2} = -4 v_t^i u^i > 0$.

To summarize, since the approximation error is decreasing for each update stage Eq. (6), this means that Eq. (19) is equivalent to Eq. (12),.e. to maximize the prediction error one should choose the next sample to be near the hypothesized separator (small $\rho$), where the distance should decrease with decreasing approximation error. This shows that maximizing the reward, given by the absolute value of the prediction error, is equivalent to active learning, i.e. the next position that maximizes the prediction error is near the current approximated plane, where the distance is bounded, with decreasing bound as the approximation becomes better. Hence, active learning indeed relates to the maximization of prediction error.

The analysis above assumed a one-step horizon from the reinforcement perspective, i.e. maximizing the next-step prediction error results in active-learning like policy. Generally, RL aims at maximizing the *accumulated* and discounted future rewards, and not a one time step reward. Thus, the converged actor of the ReAL algorithm results in an even better policy than the one presented in Eq. (19).

### 3.4. Neuronal network model

We present a biologically plausible neuronal network that implements and connects the two models, namely, perceptron based active learning (Sections 2.3 and 3.1) and reinforcement active learning (Sections 2.4 and 3.2), Fig. 3.

#### 3.4.1. Neuronal model for active learning

The model consists of two components, Fig. 3: one chooses the next action to perform (Fig. 3(black)) and the other compares the expected outcome (touch/no touch) to the actual sensory input (Fig. 3(blue)). The choice of action ($a$) is affected by the current whisker angle ($\theta$), a constant representing the neuronal threshold ($-1$) and adaptively thresholded noise ($d$). These inputs are modulated by synaptic weights, $z_\theta, z_b, z_d$, respectively, that change during learning. In turn, the current action and angle serve to predict the expected outcome ($\widehat{B}$). This is modulated by another set of learned synaptic weights ($v_\theta$, $v_a$, $v_b$). The expected outcome is compared to the actual sensory input ($B$), and generates a prediction error ($\epsilon$). Following each palpation motion, synaptic weights throughout the network are updated according to this error. The comparisons are also aggregated (Fig. 3(green)) and used to inhibit the adaptive threshold neuron ($d$).

More specifically, the implementation of the learning algorithm is straightforward, as it is perceptron based, Fig. 3(blue arrows). Eq. (6) is implemented as Hebbian learning that occurs only on a misclassification. The latter is calculated by comparing the predicted and the correct classification, Fig. 3(red arrows).

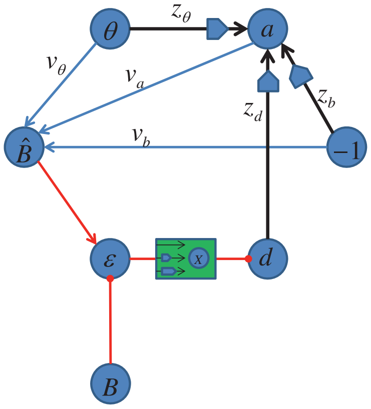

**Fig. 3.** Neuronal network implemented active learning model. Nodes represent variables and arrows represent weights. Blue arrows indicate prediction network; red arrows indicate error network and black arrows indicate action network. Arrows/round-tips indicate excitatory/inhibitory connections; blue box on top of arrows indicate delay lines, where the size of the box indicate the amount of delay; green box include $R = 3$ delay lines and a circled $X$ indicate a thresholded integrate-and-fire neuron, i.e. an AND neuron.

A unique combination of temporal integration, integrate-and-fire and accumulated inhibition implements the adaptive threshold mechanism, Fig. 3(green box). More precisely, feeding the classification error into multiple delay lines (Fig. 3(green box, black arrows)) allows the integration of classification success information over time. $R$ delay lines impinge on an integrate-and-fire neuron such that only if all inputs fire, the target neuron fires (an AND neuron) (Fig. 3(green box, circled X)); this enables the detection of $R$ consecutive correct classifications. The output of this neuron inhibits a noisy neuron, $d$, in an asymmetric accumulative fashion, i.e. it inhibits $d$ when it fires, but does excites, or ds-inhibits when silent. This exemplifies a mechanism of an adaptive decreasing threshold. In other words, only when $R$ "delayed" correct classifications coincide, the noisy neuron is inhibited and reduces its influence on the action neuron. We speculate that the function of $d$ can be implemented by 5-HT neurons, whose ability to modulate whisking had been demonstrated (Hattox et al., 2003; Harish and Golomb, 2010).

The action-based active learning and velocity dependent control results in a unique determination of the next action, Fig. 3(black). The following relations can be derived from Eqs. (1) and (12):

$$a_{t+1} = z_\theta \theta_t + z_d d - z_b \tag{20}$$

$$z_\theta = -v_\theta/(v_\theta + v_a) \quad z_d = 1 \quad z_b = -v_b/(v_\theta + v_a) \tag{21}$$

Thus, the next action is determined according to the current angle and the adaptive threshold. More precisely, the next action is chosen such that the overall state is near the "predicted" separator; however, this requires an inverse model, i.e. given the required angle, what is the action to take. The inverse model that directs action towards the predicted separator is exemplified by $z_\theta, z_b$. On top of this, the adaptive threshold modulated noise is added to allow exploration, via the $d$ neuron.

The entire network works as follows, Fig. 3: classification is predicted according to the "blue" network; it is compared to the real classification; if there is a misclassification, the "blue" and "black" connections are updated according to Hebbian learning rules and Eq. (21), respectively (see below); if the classification is correct $R$ consecutive times, the noise neuron is inhibited via the "green" network; the next action is selected according to the "black" (delayed) network.

The mechanism described above works on multiple dynamical regimes, related to the sensory-motor cycles (not whisking cycles). The first is the sub-cycle regime, which determines the dynamics of the whisker angle-change after implementing a given action (Simony et al., 2010); this regime lies in the 3–7 ms domain. The second is the cycle dynamical regime, meaning the influence of one action on the next one via an intermediate touch event, or conversely, the perception of one touch event on the perception of the next, via an intermediate action. This regime of a single sensory-motor loop lies in the range of 15–20 ms (Mitchinson et al., 2007; Deutsch et al., 2010; Nguyen and Kleinfeld, 2005). The last dynamical regime is across-cycle, in which consecutive classifications accumulate and change the adaptive threshold; it lies in the 50–200 ms domain.

#### 3.4.2. Neuronal model for ReAL

The reinforcement active learning neuronal implementation is also straightforward (Doya, 2007; Schultz, 2010). The learner is identical to the perceptron-based network. The critic is a value approximator that can be modeled using radial-basis neural network that updates its weights according to the TD-error. The actor, which is a neural controller, can be modeled with a feed-forward neuronal network that also updates its weight according to the same TD-error. These neuronal models have a well established biological analogues.

### 3.4.3. Common neuronal characteristics

The common characteristic of the perceptron network and the ReAL model is that the connectivities of the action networks, Fig. 3(black), are modified according to the prediction error. This relation was presented in the previous section and its mechanistic ramifications are novel. In the perceptron network it means that the prediction error not only changes, via Hebbian learning, the perceptual network $v$ but also the action network, $z$, in a cycle-by-cycle basis. This requires a non-trivial learning mechanism in which the error signal in one network affects another network's weights. In the ReAL implementation, the prediction-error action learning means that the reward system, instantiated by the critic, receives the percepetual prediction error as an *input*; in turn it sends the TD-error to modify the actor's weights.

## 4. Discussion

The active learning field has mainly focused on labeling streams of data, wherein the cost of labeling is high (Settles, 2009). Applications to real sensory modalities have focused on the visual system, in which studies have shown that primates and humans indeed utilize policies with principles based on active learning (Itti and Koch, 2000). However, analytical and numerical analysis of optimal active touch has been scant.

On the other hand, a new emergent field is that of intrinsic reward RL (Barto et al., 2004; Oudeyer et al., 2007), also known as theory of creativity (Schmidhuber, 2010), and is widely used in developmental robotics (Weng, 2004). This field assumes that rewards come from learning, usually the predictions of state/action/reward relations. A new typology is suggested in (Schmidhuber, 2010), where each implementation should describe the predictors, the intrinsic reward and the RL controllers.

However, most of the applications focused on high-level cognitive and behavioral aspects (Weng, 2004; Oudeyer et al., 2007; Schmidhuber, 2010). Here, we applied the ReAL paradigm to low level, sensory-motor space which consists of the whisker angle and velocity. Furthermore, previous examples focused on visual inputs (Weng, 2004; Oudeyer et al., 2007), whereas applications of intrinsic reward to the tactile domain are few (Gordon and Ahissar, 2011).

The active learning algorithm is based on a perceptron learning rule and as such has intrinsic biologically-oriented reasoning and can be easily implemented using neuronal network models. The unique feature of such networks are the *combined* learning of both the predictor- and the action-networks via a single common signal, namely, the prediction error. The prediction network changes according to the ubiquitous Hebbian learning rule, whereas the action determining network has a more complex learning rule. Nevertheless, these learning networks are tightly connected.

Furthermore, the active learning algorithm mandates an adaptive threshold whose strength relates to the cumulative prediction successes. Such a mechanism was suggested above that combines delay lines and integrate-and-fire neurons, both abundant in the central nervous system (Dan and Poo, 2006; Sjstrm et al., 2003). Another important ingredient is the "noise generator" that becomes inhibited as success increases. Noise generation is thought to occur in the basal ganglia and is crucial for explorative behavior (Tumer and Brainard, 2007; Kao et al., 2008; Sober et al., 2008; Andalman and Fee, 2009), such as the one considered here.

However, a crucial mechanism in the adaptive threshold is the asymmetric cumulative learning, i.e. the fact that the noise is diminished in an across-cycle dynamical regime; it does not increase back after it was inhibited as long as the perceptual task continues. For perceptual processes in the order of a few hundreds of ms, mechanisms involving $\text{GABA}_\text{B}$ inhibition (Golomb et al., 2006) and/or neuromodulatory modulation could apply. As 5-HT involvement in the modulation of whisking had been indicated (Hattox et al., 2003; Harish and Golomb, 2010), it becomes a leading candidate for the implementation of this mechanism.

In the context of the perceptual task of object localization, one must also define within the model when the task ends (Ahissar and Knutsen, 2008; Knutsen and Ahissar, 2009; Horev et al., 2011). In the perceptron based model, we hypothesize that the decreased noise generation plays the role of perceptual confidence, i.e. when the explorative factor of the model decreases below a certain threshold, the agent perceives the object with enough confidence and the task ends. Combined with the hypothesis that the mechanism of the decreased adaptive threshold is mediated via neuromodulators (Ahissar et al., 1996; Bahar et al., 2004), e.g. 5-HT in this case, it raises the prediction that 5-HT may correlate with perceptual confidence. More precisely, that the levels of 5-HT in the relevant areas change monotonically as the task progresses and reaches a certain level when the animal "reports" its decision, e.g. goes to the correct sipper (Knutsen and Ahissar, 2009; Horev et al., 2011). This speculation is in line with previous suggestions (Friston et al., 1991; Rogers et al., 1999).

The intrinsic-reward reinforcement learning model describes the emergence of behavior and thus promotes mainly developmental predictions, i.e. behaviors and their underlying neural circuitry during the critical period of development in pups. One straightforward prediction is that pups do not palpate novel objects immediately, i.e. once their whiskers are grown enough to reach objects, the model predicts that their first behavior should be quasi-random. This prediction has been recently supported in Grant et al. (2011).

Furthermore, the converged behaviors are strongly dependent on the experience of the pup, thus changing pups' experience should produce different emergent behaviors. For example, partially paralyzing the mystacial pad muscles during development, i.e. reducing their responsiveness and contracting strength, should result in a different object palpation when they are adults, even if at adulthood there is no paralysis. Similarly, affecting the sensory input during development, e.g. via pharmacological manipulations along the sensory pathway, should result in markedly different behaviors in adulthood. Furthermore, preventing whisker-object touch during development, e.g. by attaching plastic cones to the snout, should result in the lack of palpation behavior during adulthood. Another option is to place the entire home-cage in a puff-ball material, such that the pups never encounter a hard, inflexible objects. In such a manner, the palpation behavior observed in normal rats should be drastically changed.

In order to verify these predictions, monitoring their whisking behavior is mandatory along the developmental axis. Two options are possible, determined on the available technology and design of the home cage. The first is continuous monitoring of the pup *within* the home cage during its development. This option is very difficult due to the stringent requirements of the specialized video and tracking system of the whiskers. The second option is to extract the pups from their home cage in a relatively high frequency and monitor them in a controlled environment. This type of analysis of pup whisker behavior is in its infancy (Grant et al., 2011), but the rapid advancement of tracking techniques can expedite the performance of the proposed experiments.

The RL model also predicts novel neural circuitry during development. In order to facilitate rewarding prediction error, there should be a strong *input* connectivity to the rewarding system from internal model areas, e.g. barrel cortex. The model predicts that this connectivity should be stronger during development to allow convergence of the stereotypical whisking behaviors apparent in adult rats. Furthermore, the conveyed information in these connections should code error signals (Shadmehr and Krakauer, 2008; Lalazar and Vaadia, 2008). The anatomical and functional circuitry of the developing pup is mostly unknown, yet the underlying infrastructure for the proposed curiosity loops should be evident to corroborate the proposed model.

## 5. Conclusion

Rats palpate novel objects with their whiskers in order to ascertain their location, shape and texture. This behavior was reproduced using two compatible approaches, namely, active learning and reinforcement learning with intrinsic reward. We have modified a perceptron-based active learning algorithm (Dasgupta et al., 2009) to accommodate membership queries (Settles, 2009), bounded space and velocity-dependent control. Our analysis shows that object palpation, i.e. alternating between touch and no touch, learns the proper linear separator exponentially faster than random actions.

Additionally, we have augmented the incremental Natrual Actor Critic (Bhatnagar et al., 2007) reinforcement learning algorithm with reward that is equal to the square of the prediction error of learning the forward model of touch (Gordon and Ahissar, 2011). This resulted in a remarkably similar behavior to that of the aforementioned active learning algorithm. We have then showed that indeed the two approaches are tightly connected.

The two models presented and their resultant behavior, that is highly reminiscent of rats palpation behavior, suggest that rats employ active learning and/or intrinsic reward principles when exploring their environment. Further comparison of rat and model behavior may reveal the dependency of update rules on correlation of neuronal activities, behavioral factors and modulatory neuronal systems (Ahissar et al., 1998; Ego-Stengel et al., 2001; Schultz, 2010).

Furthermore, the ReAL model suggests that such behavior is learned, rather than innate (Weng, 2004), and is experience-dependent. It thus predicts that pups' initial contact-induced whisker behavior would be distinctively different and much more random than the structured palpation seen in adult rats.

Finally, developmental robotics (Weng, 2004; Asada et al., 2009) can benefit from both proposed models as there are recent attempts to employ artificial whisker-based sensors (Solomon and Hartmann, 2006; Evans et al., 2010; Sullivan et al., 2012). Implementing ReAL concepts in such robots may make them much more efficient and autonomous.

## Acknowledgments

This work was supported by EU Grant BIOTACT (ICT-215910), the Israeli Science Foundation Grant No. 749/10, the United States-Israel Bi-national Science Foundation (BSF) Grant No. 2007121, and the Minerva Foundation funded by the Federal German Ministry for Education and Research. G.G. was supported by the Clore fellowship. E.A. holds the Helen Diller Family Professorial Chair of Neurobiology.

## References

Adejumo, A., Engelbrecht, A.P., 1999. A comparative study of neural network active learning algorithms. In: International Conference on Artificial Intelligence, pp. 32–35).

Ahissar, E., Abeles, M., Ahissar, M., Haidarliu, S., Vaadia, E., 1998. Hebbian-like functional plasticity in the auditory cortex of the behaving monkey. Neuropharmacology 37, 633–655.

Ahissar, E., Haidarliu, S., Shulz, D.E., 1996. Possible involvement of neuromodulatory systems in cortical hebbian-like plasticity. J. Physiol. Paris 90, 353–360.

Ahissar, E., Knutsen, P.M., 2008. Object localization with whiskers. Biol. Cybern. 98, 449–458.

Andalman, A.S., Fee, M.S., 2009. A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors. Proc. Natl. Acad. Sci. 106, 12518–12523.

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., Yoshida, C., 2009. Cognitive developmental robotics: a survey. Auton. Mental Dev., IEEE Trans 1, 12–34.

Bahar, A., Dudai, Y., Ahissar, E., 2004. Neural signature of taste familiarity in the gustatory cortex of the freely behaving rat. J. Neurophysiol. 92, 3298–3308.

Barto, A.G., Singh, S., Chentanez, N., 2004. Intrinsically motivated learning of hierarchical collections of skills. In: International Conference on Developmental Learning (ICDL).

Berg, R.W., Kleinfeld, D., 2003. Rhythmic whisking by rat: retraction as well as protraction of the vibrissae is under active muscular control. J. Neurophysiol. 89, 104–117.

Bhatnagar, S., Sutton, R., Ghavamzadeh, M., Lee, M., 2007. Incremental natural actor-critic algorithms. In: Twenty-First Annual Conference on Advances in Neural Information Processing Systems, pp. 105–112.

Birdwell, J.A., Solomon, J.H., Thajchayapong, M., Taylor, M.A., Cheely, M., Towal, R.B., Conradt, J., Hartmann, M.J.Z., 2007. Biomechanical models for radial distance determination by the rat vibrissal system. J. Neurophysiol. 98, 2439–2455.

Dan, Y., Poo, M.-M., 2006. Spike timing-dependent plasticity: from synapse to perception. Physiol. Rev. 86, 1033–1048.

Dasgupta, S., Hsu, D., 2008. Hierarchical sampling for active learning. In: Proceedings of the 25th International Conference on Machine learning, vol. 1390183. ACM, pp. 208–215.

Dasgupta, S., Kalai, A., Monteleoni, C., 2009. Analysis of perceptron-based active learning. J. Mach. Learn. Res. 10, 281–299.

Deutsch, D., Pietr, M., Knutsen, P., Ahissar, E., Schneidman, E., 2010. Feedback After Touch Modifies Rat's Whisking.

Doya, K., 2007. Reinforcement learning: computational theory and biological mechanisms. HFSP J. 1, 30–40.

Ego-Stengel, V., Shulz, D.E., Haidarliu, S., Sosnik, R., Ahissar, E., 2001. Acetylcholine-dependent induction and expression of functional plasticity in the barrel cortex of the adult rat. J. Neurophysiol. 86, 422–437.

Evans, M., Fox, C., Pearson, M., Lepora, N., Prescott, T., 2010. Whisker-object contact speed affects radial distance estimation. In: IEEE International Conference on Robotics and Biomimetics (ROBIO).

Friston, K.J., Grasby, P.M., Frith, C.D., Bench, C.J., Dolan, R.J., Cowen, P.J., Liddle, P.F., Frackowiak, R.S., 1991. The neurotransmitter basis of cognition: psychopharmacological activation studies using positron emission tomography. CIBA Found Symp. 163, 76–87, discussion 87–92.

Gao, P., Bermejo, R., Zeigler, H.P., 2001. Whisker deafferentation and rodent whisking patterns: behavioral evidence for a central pattern generator. J. Neurosci. 21, 5374–5380.

Golomb, D., Ahissar, E., Kleinfeld, D., 2006. Coding of stimulus frequency by latency in thalamic networks through the interplay of gabab-mediated feedback and stimulus shape. J. Neurophysiol. 95, 1735–1750.

Gordon, G., Ahissar, E., 2011. Reinforcement active learning hierarchical loops. In: International Joint Conference on Neural Networks (IJCNN).

Gordon, G., Ahissar, E., 2012. Hierarchical curiosity loops and active sensing. Neural Networks.

Govindhasamy, J.J., McLoone, S.F., Irwin, G.W., French, J.J., Doyle, R.P., 2005. Reinforcement learning for online control and optimisation. In: IEE Control Engineering Book Series, vol. 70, pp. 293–326 (Chapter 9).

Grant, R.A., Mitchinson, B., Fox, C.W., Prescott, T.J., 2009. Active touch sensing in the rat: anticipatory and regulatory control of whisker movements during surface exploration. J. Neurophysiol. 101, 862–874.

Grant, R.A., Mitchinson, B., Prescott, T.J., 2011. The development of whisker control in rats in relation to locomotion. Dev. Psychobiol.

Harish, O., Golomb, D., 2010. Control of the firing patterns of vibrissa motoneurons by modulatory and phasic synaptic inputs: a modeling study. J. Neurophysiol. 103, 2684–2699.

Hattox, A., Li, Y., Keller, A., 2003. Serotonin regulates rhythmic whisking. Neuron 39, 343–352.

Horev, G., Saig, A., Knutsen, P.M., Pietr, M., Yu, C., Ahissar, E., 2011. Motor-sensory convergence in object localization: a comparative study in rats and humans. Philos. Trans. R Soc. Lond. B Biol. Sci. 366, 3070–3076.

Itti, L., Koch, C., 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res. 40, 1486–1489.

Jordan, M.I., 1992. Forward models: supervised learning with a distal teacher. Cognitive Sci. 16, 307–354.

Kao, M.H., Wright, B.D., Doupe, A.J., 2008. Neurons in a forebrain nucleus required for vocal plasticity rapidly switch between precise firing and variable bursting depending on social context. J. Neurosci. 28, 13232–13247.

Knutsen, P.M., Ahissar, E., 2009. Orthogonal coding of object location. Trends Neurosci. 32, 101–109.

Knutsen, P.M., Biess, A., Ahissar, E., 2008. Vibrissal kinematics in 3d: tight coupling of azimuth, elevation, and torsion across different whisking modes. Neuron 59, 35–42.

Kolodziejski, C., Porr, B., Wrgtter, F., 2009. On the asymptotic equivalence between differential hebbian and temporal difference learning. Neural Comput. 21, 1173–1202.

Lalazar, H., Vaadia, E., 2008. Neural basis of sensorimotor learning: modifying internal models. Curr. Opin. Neurobiol. 18, 573–581.

Mitchinson, B., Martin, C.J., Grant, R.A., Prescott, T.J., 2007. Feedback control in active sensing: rat exploratory whisking is modulated by environmental contact. Proc. Biol. Sci. 274, 1035–1041.

Nguyen, Q.T., Kleinfeld, D., 2005. Positive feedback in a brainstem tactile sensorimotor loop. Neuron 45, 447–457.

Oudeyer, P.Y., Kaplan, F., Hafner, V.V., 2007. Intrinsic motivation systems for autonomous mental development. IEEE Trans. Evol. Comput. 11, 265–286.

Rogers, R.D., Everitt, B.J., Baldacchino, A., Blackshaw, A.J., Swainson, R., Wynne, K., Baker, N.B., Hunter, J., Carthy, T., Booker, E., London, M., Deakin, J.F., Sahakian, B.J., Robbins, T.W., 1999. Dissociable deficits in the decision-making cognition of chronic amphetamine abusers, opiate abusers, patients with focal damage to prefrontal cortex, and tryptophan-depleted normal volunteers: evidence for monoaminergic mechanisms. Neuropsychopharmacology 20, 322–339.

Schmidhuber, J., 1990. A possibility for implementing curiosity and boredom in model-building neural controllers. Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats, 116542. MIT Press, pp. 222–227.

Schmidhuber, J., 2010. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Trans. Auton. Mental Dev. 2, 230–247.

Schultz, W., 2010. Dopamine signals for reward value and risk: basic and recent data. Behav. Brain Funct. 6, 24.

Settles, B., 2009. Active Learning Literature Survey. Computer Sciences Technical Report University of Wisconsin Madison.

Shadmehr, R., Krakauer, J.W., 2008. A computational neuroanatomy for motor control. Exp. Brain Res. 185, 359–381.

Simony, E., Bagdasarian, K., Herfst, L., Brecht, M., Ahissar, E., Golomb, D., 2010. Temporal and spatial characteristics of vibrissa responses to motor commands. J. Neurosci. 30, 8935–8952.

Sjstrm, P.J., Turrigiano, G.G., Nelson, S.B., 2003. Neocortical ltd via coincident activation of presynaptic NMDA and cannabinoid receptors. Neuron 39, 641–654.

Sober, S.J., Wohlgemuth, M.J., Brainard, M.S., 2008. Central contributions to acoustic variation in birdsong. J. Neurosci. 28, 10370–10379.

Solomon, J.H., Hartmann, M.J., 2006. Biomechanics: robotic whiskers used to sense features. Nature 443, 525.

Sullivan, J.C., Mitchinson, B., Pearson, M.J., Evans, M., Lepora, N.F., Fox, C.W., Melhuish, C., Prescott, T.J., 2012. Tactile discrimination using active whisker sensors. IEEE Sensors J. 12, 350–362.

Tumer, E.C., Brainard, M.S., 2007. Performance variability enables adaptive plasticity of /‘crystallized/’ adult birdsong. Nature 450, 1240–1244.

Weng, J., 2004. Developmental robotics: theory and experiments. Int. J. Humanoid Robotics 1, 199–236.