J Comput Neurosci (2014) 37:259–280
DOI 10.1007/s10827-014-0500-1

Learning and control of exploration primitives

Goren Gordon · Ehud Fonio · Ehud Ahissar

Received: 25 July 2013 / Revised: 11 February 2014 / Accepted: 12 March 2014 / Published online: 7 May 2014
© Springer Science+Business Media New York 2014

Abstract Animals explore novel environments in a cautious manner, exhibiting alternation between curiosity-driven behavior and retreats. We present a detailed formal framework for exploration behavior, which generates behavior that maintains a constant level of novelty. Similar to other types of complex behaviors, the resulting exploratory behavior is composed of exploration motor primitives. These primitives can be learned during a developmental period, wherein the agent experiences repeated interactions with environments that share common traits, thus allowing transference of motor learning to novel environments. The emergence of exploration motor primitives is the result of reinforcement learning in which information gain serves as intrinsic reward. Furthermore, actors and critics are local and ego-centric, thus enabling transference to other environments. Novelty control, i.e. the principle which governs the maintenance of constant novelty, is implemented by a central action-selection mechanism, which switches between the emergent exploration primitives and a retreat policy, based on the currently-experienced novelty. The framework has only a few parameters, wherein time-scales, learning rates and thresholds are adaptive, and can thus be easily applied to many scenarios. We implement it by modeling the rodent's whisking system and show that it can explain characteristic observed behaviors. A detailed discussion of the framework's merits and flaws, as compared to other related models, concludes the paper.

Keywords Reinforcement learning · Intrinsic reward · Bayesian inference · Information gain · Hierarchical control · Motor primitives

Action Editor: T. Sejnowski

G. Gordon (✉) · E. Ahissar
Adaptive Perceptual Processing Lab, Weizmann Institute of Science, Department of Neuro biology, Rehovot , Israel
e-mail: goren@gorengordon.com

E. Ahissar
e-mail: ehud.ahissar@weizmann.ac.il

E. Fonio
Ant Collective Behavior Group, Weizmann Institute of Science, Department of Physics of Complex Systems, Rehovot, Israel
e-mail: ehud.fonio@weizmann.ac.il

1 Introduction

Animals and humans interact and explore their surrounding environment by moving. They exhibit complex behavioral patterns motivated by both curiosity and fear of novel sensations (Tinbergen 1951; Barnett 1958; Misslin and Cigrang 1986; File 2001; Elliot 2006; Hughes 2007; Fonio et al. 2009). The goal of the proposed formal framework is to capture the intricate interaction between the curiosity drive and fear in exploring animals (Barnett 1958; Misslin and Cigrang 1986; Fonio et al. 2009). This neurophysiologically-plausible formal framework aims at maintaining a *constant* level of novelty and is composed of discrete, learnable *exploration motor primitives*. The framework builds upon and extends the model presented in (Gordon and Ahissar 2012) and is in contrast to previous models that suggested maximizing or minimizing novelty (in its many forms, see Little and Sommer 2013).

We present a framework that describes the emergence of these exploration motor primitives from intrinsic motivation principles and their control strategy, Fig. 1. The primitives are *local* policies, learned through intrinsic reward reinforcement learning (RL), and are aimed at learning specific sensory-motor correlations. Due to their unique local nature, they are transferable to novel environments that share similar traits (Konidaris and Barto 2007). The exploration primitives emerge in a sequence, where each new primitive focuses on a different novel aspect of the agent-environment interaction. The motor strategy is composed of switching between these exploration primitives and a retreat primitive, where the latter is an adaptive optimal behavior aimed at reducing novelty in a safe manner (Moldovan and Abbeel 2012). The retreat policy exemplifies the notion of fear, yet instead of incorporating a new intrinsic source of reward/punishment, we link it to the accumulation of too much novelty, i.e. neophobia. In this formulation, a balance between the curiosity drive (neophilia) and fear (neophobia) is maintained by a single variable, namely, novelty. The emerged complex behavior is a structured alternation between exploration and retreats, in accordance with approach-avoidance (Tinbergen 1951; Barnett 1958; Misslin and Cigrang 1986; File 2001; Elliot 2006; Hughes 2007; Fonio et al. 2009).

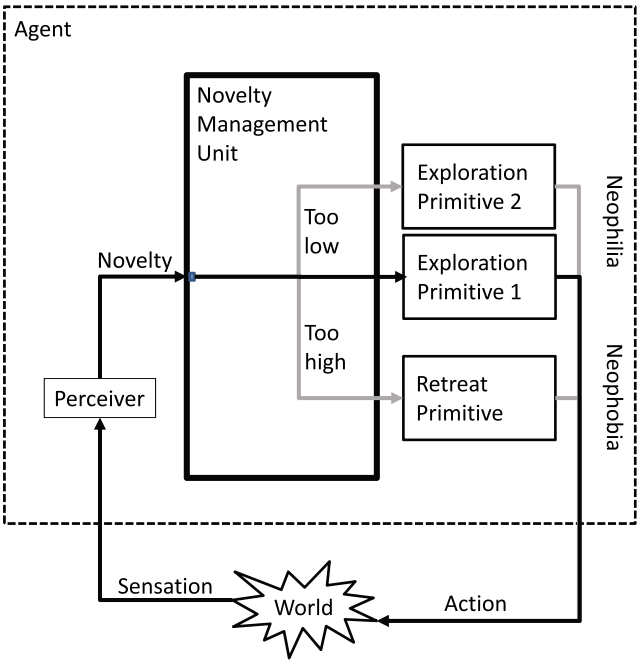

Fig. 1 Framework architecture. The agent perceives the world by attempting to predict its sensations. Errors in prediction result in novelty, which is managed via a balance between neophobia, i.e. too much novelty initiates retreat, and neophilia, i.e. too little novelty activates emergent exploration motor primitives

2 Related work

Animals employ different motor strategies that depend on the specific context and desired goal. It was shown, however, that there are fundamental concepts that guide these motor strategies. Complex movements were shown to be composed of motor building blocks (Richardson and Flash 2002; Flash and Hochner 2005) which can be combined using specific syntactic rules. However, most of the analyzed motions were concerned with motor-guided motor control, i.e. the objective was motion, such as reaching or grasping (Flash and Handzel 2007). A different type of motor strategy appears when the goal is sensation, also known as active sensing (Szwed et al. 2003; Kleinfeld et al. 2006; Gordon et al. 2011).

In this contribution we are concerned with motor control learned via reinforcement learning (RL) (Kaelbling et al. 1996; Schultz et al. 1997; Fox et al. 2008), wherein the goal is to maximize future accumulated rewards, attained by adapting policies. Building a complex motor strategy that maximizes rewards has been investigated in the context of Hierarchical RL (Barto and Mahadevan 2003) and in the options setup (Sutton et al. 1999; Stolle and Precup 2002), where the task is divided to sub-tasks, each with its optimal policy.

However, in most experimental setups and robotics scenarios, the task, goal or reward are defined extrinsically, e.g., the experimenter decides the task and grants rewards according to its performance. In scenarios considered here, namely, exploration of novel environments, there is no external reward or a pre-defined goal-state, but rather there is an intrinsic motivation to learn about the environment in the most efficient manner. Such a curious agent, whether an animal, human or robot, interacts with its environment in order to learn cause and effect relations. The formalization of curiosity in this context is the notion of rewarding novelty, or surprise, such that actions that increase the information gain of the agent are rewarded and hence become more likely to recur. The concept of rewarding novelty is also known as theory of creativity (Schmidhuber 2010), and is widely used in developmental robotics (Weng 2004; Oudeyer et al. 2007). Furthermore, several information-theoretic approaches to minimize surprise (Friston 2010), maximize novelty (Tishby and Polani 2011; Little and Sommer 2013) or maximize controllability (Polani 2009) of the agent have been recently developed.

The interaction between the curiosity drive and other primary rewards was less analyzed (Vergassola et al. 2007; Tishby and Polani 2011). An information-theoretic perspective in (Tishby and Polani 2011) attempts to relate external reward and the information required to acquire it. It derives an "info-Bellman" equation where information serves as a symmetric counter-part to rewards from external sources.

However, to the best of our knowledge, fear and anxiety have not been incorporated into these types of models, even though they are ubiquitous in nature (Barnett 1958; Misslin and Cigrang 1986; File 2001; Whishaw et al. 2006; Hughes 2007).

One active sensing system often studied is the rodent's whisker system (Kleinfeld et al. 2006). It was recently shown that initial contacts of rodents' whiskers with novel objects were immediately followed by rapid retraction (Mitchinson et al. 2007). Furthermore, repeated contacts with the same object showed increased complexity of contact, termed "touch-induced-pump" (Deutsch et al. 2012). We hypothesize that the aforementioned behaviors are composed of exploration motor primitives, interleaved with retreat primitives. We model this system using the proposed framework and analyze the emergent behavior in view of the aforementioned observed behaviors.

3 Reinforcement learning background

In the framework presented below, we use the reinforcement learning (RL) paradigm, more specifically, the incremental Natural Actor Critic (iNAC) algorithm (Bhatnagar et al. 2007), which is an efficient actor-critic-based reinforcement learning algorithm (see (Gordon and Ahissar 2012) for more details). The agent selects an action, $\mathbf{a}_t$, at each time t using a randomized stationary policy, designated as the actor: $\pi(\mathbf{a}|\mathbf{s}) = \Pr(\mathbf{a}_t = \mathbf{a}|\mathbf{s}_t = \mathbf{s}; \lambda_t)$, where λ_t are the actor parameters to be tuned. The Natural Actor-Critic algorithm uses the compatible functions, defined as $\psi(\mathbf{s}_t, \mathbf{a}_t) = \nabla_\lambda \pi(\mathbf{a}_t|\mathbf{s}_t)$, i.e. a set of functions that represents how the actor depends on each tuning parameter, λ. The critic, $\hat{V}^\pi(\mathbf{s}; \nu_t)$, is the approximation of the true value function, $V^\pi(\mathbf{s})$, which measures the expected future accumulated rewards for an initial state $\mathbf{s}$ and behaving according to the policy π. The critic thus attempts to learn this value function by tuning the parameters ν_t using the following set of functions $\phi(\mathbf{s}_t) = \nabla_\nu \hat{V}^\pi(\mathbf{s}; \nu_t)$, which similar to the compatible functions, represent how the critic depends on each tuning parameter.

Using the critic, the reinforcement learning algorithm computes the temporal difference (TD) error (Sutton 1988), here taken to be:

$$\delta_t = r_t - \hat{J}_{t+1} + \hat{V}^\pi(s_t; \nu_t) - \hat{V}^\pi(s_{t+1}; \nu_t), \tag{1}$$

where $\hat{J}_t$ is the estimated average reward. The TD-error, δ_t, represents the error in the critic's prediction of the value function, compared to the actual received rewards, normalized to the estimated average reward. In other words, the TD-error measures the difference between what the critic "thought" the reward should be ($\hat{V}^\pi(s_{t+1}) - \hat{V}^\pi(s_t)$), i.e. the change in value due to change in states; and the actual (normalized) reward ($r_t - \hat{J}_{t+1}$). If the critic approximates the value function well, it should accurately predict the reward and the TD-error should be small; conversely, if the critic does not differentiate values at all ($\hat{V}^\pi(\mathbf{s}) = \hat{V}^\pi \ \forall \mathbf{s}$), the TD-error is proportional to the (normalized) reward. The TD-error is thus used to update the tuning parameters of the critic in order to better approximate the value function; it is also used to update the actor's tuning parameters to have a better policy that will maximize the future accumulated rewards.

Our version of the iNAC algorithm thus have several parameters that are updated. The update rules are summarized below (for more details, see Gordon and Ahissar 2012):

$$\hat{J}_{t+1} = (1 - \xi)\hat{J}_t + \xi r_t \tag{2}$$

$$\nu_{t+1} = \nu_t + \alpha^C \delta_t \phi(s_t) \tag{3}$$

$$w_{t+1} = \left[I - \alpha^C \psi(s_t, a_t)\psi(s_t, a_t)^T \right] w_t + \alpha^C \delta_t \psi(s_t, a_t) \tag{4}$$

$$\lambda_{t+1} = \lambda_t + \alpha^A w_{t+1} \tag{5}$$

where ξ is the average reward update rate, w_t are the advantage parameters, α^C, α^A are the learning rates of the critic and actor, respectively.

4 Formal framework of exploration behavior

The framework presented here is composed of three main components: (i) a novelty-seeking component, composed of a set of hierarchical curiosity loops, whereby the agent converges to behaviors that optimize learning; (ii) a novelty-aversive component composed of a single motor-primitive dubbed retreat and; (iii) a novelty management unit that controls the transition between these motor primitives. We first present the basic curiosity loop, which is composed of the exploration perceiver and an actor-critic module. We then describe the hierarchical buildup of curiosity loops. The introduction of the retreat motor-primitive follows; this behavior, which exemplifies the notion of neophobia, aims at reducing novelty by exploiting information already learned in order to return to the most familiar base-state of the system. We then present the principles that govern the novelty management unit and the transition between the loops, namely, "novelty seeking" that determines advancement to higher loops and "novelty avoidance" that determines the transition to the retreat policy. A summary of the framework parameters is followed by a description of the environmental settings.

4.1 Novelty-seeking: the basic curiosity loop

Each curiosity loop attempts to find the policy, or Actor, that optimizes learning the sensory-motor correlations of the animal's interaction with its environment, represented by the perceiver module. The curiosity loops are thus constructed from perceivers, critics and actors.

Exploration perceiver The perceiver attempts to correctly predict the future sensory information given the current sensory state and the applied actions or motor commands; this is also known as the forward model (Ouyang et al. 2006; Behera et al. 1995; Shadmehr and Krakauer 2008; Lalazar and Vaadia 2008; Kawato 1999). The perceiver is thus formulated as

$$L(\mathbf{s}'|\mathbf{s},\mathbf{a}) = \Pr(\mathbf{s}_{t+1} = \mathbf{s}'|\mathbf{s}_t = \mathbf{s}, \mathbf{a}_t = \mathbf{a}) \tag{6}$$

where boldface letters denote vectors; $\mathbf{s}_t$ is the sensory state (vector) at time t; $\mathbf{a}_t$ is the action (vector) performed at time t; and $\mathbf{s}_{t+1}$ is the predicted next sensory state (vector).

The state space can be composed of all sensory information, either proprioceptive regarding the agent's own body, its location in space or its input from its sensory organs. The action space represents motor commands to muscles, movement in an arena, etc. For brevity, we define the perceiver as:

$$P(\mathbf{s}') \equiv L(\mathbf{s}'|\mathbf{s},\mathbf{a}) \tag{7}$$

i.e. the conditional probability function of the next sensory state, $\mathbf{s}'$, given the current state and action. At each time step, a state is observed by the agent, denoted by $\mathbf{o}$, which, due to noise, may be different from the actual state, $\mathbf{s}$. The perceiver probability function is updated each time step by this observation:

$$\begin{aligned} P_{t+1}(\mathbf{s}'|\mathbf{o}) &= L(\mathbf{s}'|\mathbf{s},\mathbf{a},\mathbf{o}) \\ &= \Pr(\mathbf{s}_{t+1} = \mathbf{s}'|\mathbf{s}_t = \mathbf{s}, \mathbf{a}_t = \mathbf{a}, \mathbf{o}_{t+1} = \mathbf{o}) \end{aligned} \tag{8}$$

In other words, given the new observed sensory input $\mathbf{o} = \mathbf{o}_{t+1}$, the probability function of all possible sensory inputs $\mathbf{s}$ are updated via Bayes theorem:

$$P_{t+1}(\mathbf{s}'|\mathbf{o}) = P_t(\mathbf{s}')\frac{q(\mathbf{o}|\mathbf{s}')}{\sum_{\mathbf{k}} P_t(\mathbf{k})q(\mathbf{o}|\mathbf{k})} \tag{9}$$

where $q(\mathbf{o}|\mathbf{k})$ defines the probability that given the true (actual) state $\mathbf{k}$, a specific state $\mathbf{o}$ will be observed. This probability represents the noise in the sensory system, i.e. how stochastic the actual state becomes as it is observed by the agent. The initial perceiver, $P_0(\mathbf{s}')$, i.e. the perceiver in the first time step $t = 0$, is the prior the agent has of its sensory-motor correlations. It is usually assumed to be completely random, i.e. the probabilities for each next sensory state are equal and do not depend on the currently observed states.

Information gain as intrinsic reward By performing a sequence of actions and receiving sensory information, the perceiver updates its prediction probability according to Eq. (9). The change in the perceiver can be measured using the Kullback-Leibler divergence, $D_{KL}(P_{t+1}||P_t)$, which is a known measure for the amount of useful information, or information gain, of the new observed state:

$$D_{KL}(P_{t+1}(\mathbf{s}'|\mathbf{o})||P_t(\mathbf{s}')) = \sum_{\mathbf{s}'} P_{t+1}(\mathbf{s}'|\mathbf{o}) \log\left(\frac{P_{t+1}(\mathbf{s}'|\mathbf{o})}{P_t(\mathbf{s}')}\right) \tag{10}$$

In other words, upon performing action $\mathbf{a}$ a new observation, $\mathbf{o}$ is obtained resulting in an update of the perceiver; this update's intrinsic reward is measured by the information gain. The motivation of using the information gain is that one assumes that the new updated perceiver more closely resembles the true sensory-motor correlations; it then follows from the definition of the KL–divergence, that it actually measures the number of extra bits required to describe these correlations using the previous perceiver, compared to using the new updated perceiver. In other words, it costs more bits to "code the world" using the previous (not-updated) perceiver; the new (updated) perceiver is more optimal in coding the world and the KL–divergence measures by how much it is so; it measures the gained information due to the update. This information-gain is then used as an intrinsic-reward for reinforcement learning.

The RL paradigm requires a reward function. In the curiosity loop, the reward is intrinsic, i.e. it is not supplied as an external function (Barto et al. 2004; Weng 2004; Oudeyer et al. 2007; Schmidhuber 2010), (Fig. 2). As reward, we have used the information gain due to the perceiver updates, measured by the KL–divergence, Eq. (10), such that the loop learns from its (corrected) mistakes. In other words, the learning process is proportional to the information gain, i.e. the larger the gain, the more the agent has learned about its environment. Thus, by rewarding larger gains, that come from larger updates due to larger (corrected) mistakes, the behavior is encouraged to seek places it does not know (Gordon and Ahissar 2012).

Actor-Critic module An important and novel feature in our actor-critic design is the fact that we require the critic and the actor to be *local*, i.e to be independent of size and shape of the environment and the exact location of the agent within it. The local nature of the actor-critic module enables two things: (i) encountering a novel environment with different size and shape does not require development of new motor primitives (actors), but rather they are transferable, making their use ubiquitous (Konidaris and Barto 2007; Frommberger and Wolter 2010) and; (ii) the motor primitives are insensitive to discretization of the state and action

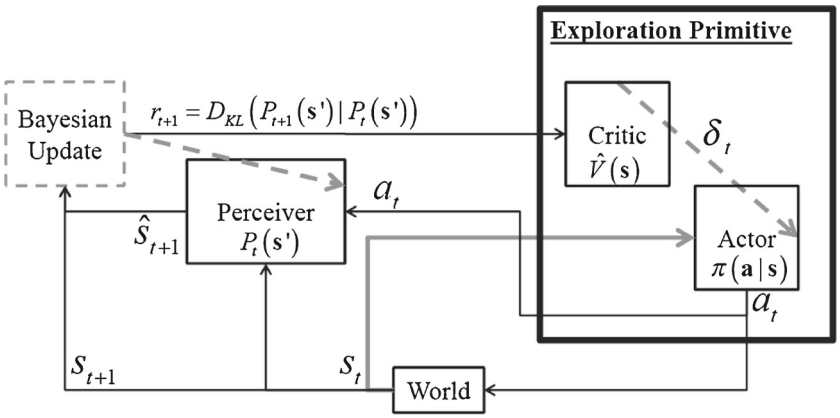

Fig. 2 The basic curiosity loop. $\mathbf{s}_t, \mathbf{s}_{t+1}, \hat{\mathbf{s}}_{t+1}$ and $\mathbf{a}_t$ are the current state, next state, predicted next state and current action, respectively. The perceiver is represented by the next-state probability function $P_t(\mathbf{s}_{t+1})$ and is updated using Bayes update, Eq. (9). The change in the perceiver, measured by the Kullback-Leibler divergence, $D_{KL}(P_{t+1}||P_t)$ (Eq. (10)), is used as an intrinsic reward for an actor-critic architecture. The Critic approximates the value of each state, $\hat{V}(\mathbf{s})$ and calculates the the TD-error δ_t, Eq. (1), which is used to update the Critic and Actor. The Actor, represented by a stochastic policy, $\pi(\mathbf{a}_t|\mathbf{s}_t)$, determines the action to be performed

spaces, but rather represent a qualitative behavior that does not change with different discretization processes.

Hence, the critic and actor do not depend on the entire state-space, but rather depend only on local aspects that are invariant to translation, rotation and discretization. We call these critics and actors "invariant", to distinguish them from other possible critics and actors that can depend on the entire state- and action-spaces. We thus separate the state, $\mathbf{s}_t$, to its invariant part, denoted by $\tilde{\mathbf{s}}_t$ and to its complementary subspace, denoted by $\check{\mathbf{s}}_t$. For example, in a scenario in which a rodent move its whiskers and touches objects, the actions are determined solely ont he basis of the sensory information of detected contact, a binary local state-space, and not on the angle of the whisker, whose discretization and extent may change over time or between environments. Thus, the same invariant actor, or motor primitive, can be used in different-sized whiskers without any modifications, in contrast to coordinate-dependent actors which must be re-learned in every new whisker configuration.

Furthermore, we require the actors to be *ego-centric*, i.e. to have an action-space that does not depend on global features. For example, moving in an arena should be dependent on actions such as forward/backward/left/right and not north/west/south/east. This requirement stems from the same principle of learning transference, i.e. the actors should be applicable to a variety of scenarios in order to be called "exploration motor primitives".

4.2 Hierarchical curiosity loops

The curiosity-loop algorithm starts with a random actor, which gradually converges into a motor primitive. However, since the actor is invariant, i.e. the actions depend only on the invariant subspace of the sensor-space, there is no guarantee that the perceiver, which includes the entire sensor space, succeeds in mapping the entire exploration space. It is more probable that only a single feature or aspect is learned while implementing the converged invariant actor.

In order to learn other aspects of the environment, a higher level curiosity loop is introduced, in which a new critic/actor pair is updated, until it too converges. Since the new loop's actor starts after the previous level's actor was already acted upon, the perceiver has changed from its initial prior and already contains the learned feature of the environment; this updated perceiver now serves as the "prior" during the behavior of the new loop. The new actor will converge to a behavior that learns new features, since the previous features are no longer novel and thus not rewarding. This process of adding new curiosity loops can be repeated, where each level's converged actor represents different motor primitives, each optimal for learning specific features in the environment (Fig. 3).

The rationale behind this hierarchical buildup of actor-critic-perceiver is that initially the most rewarding, i.e. novel, feature is learned by the converged actor. However, since the actor depends only on partial (invariant) sensors information, continuing acting upon it will not necessarily induce learning of other features. Thus, switching to a new curiosity loop produces an actor that maximizes learning the next most rewarding feature of the environment. The hierarchy can be extended recursively, until the perceiver has completely mapped the environment, producing no more rewards.

Formally, each level of the hierarchy, $l = 1, \ldots, N_H$ is described by a critic $\hat{V}^l(\tilde{\mathbf{s}})$ and an actor $\pi^l(\mathbf{a}|\tilde{\mathbf{s}})$, whereas there is a single perceiver for all loops, $L = p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$. The latter is updated at every time step, no matter which motor primitive is currently active. Hence, "going up" the

hierarchy teaches the same perceiver new features of the environment. It is important to note that only the active loop's parameters are updated at time step t, thus realizing an on-policy updating process, as opposed to off-policy (Precup et al. 2001; Wawrzynski and Pacut 2004), due to a different reward function for each level.

4.3 Novelty-aversive component: retreat policy

The retreat motor primitive is based on the assumption that the agent always starts from a base-state, with which it is most familiar and which contains no novelty (Bahar et al. 2004)(Fig. 3). For example, in a rodent's whisking scenario the base-state is the resting-angle of the whisker, considered to be the fully retracted angle. The retreat actor moves the agent to the base-state, through a route of least novelty, thus ensuring its reduction.

The retreat actor is deterministic and is implemented in the following way. Using the current perceiver, P_t, it finds the next action that will bring it closer to the base-state with the highest certainty, i.e. that the probability of a state-transition there is the highest. Hence, the retreat policy is given by:

$$\pi^0(\mathbf{a}|\check{\mathbf{s}}) = \begin{cases} 1 \;\; \mathbf{a} = \text{argmin}_{\mathbf{a}'}\Big[||\check{\mathbf{s}}' - \check{\mathbf{s}}_{\text{base}}|| \\ \quad \times \left(1 - P_t(\check{\mathbf{s}}'|\check{\mathbf{s}}, \mathbf{a}')\right)\Big] \\ 0 \;\; \text{otherwise} \end{cases} \tag{11}$$

In other words, it looks for a neighboring state, that it can reach with one action, which can be reached with certainty, e.g. no intervening obstacles, that reduces the distance between the current position and the base-state. In addition, it remembers the followed path to the base-state and disregards actions that lead to a repeated trajectory, i.e. it avoids circular paths. If the actor is continuously active, the agent will follow the most certain path to the base-state.

4.4 Novelty controller

Novelty is represented in the framework by the information gain, which also serves as the reward. The novelty management unit determines the transition between the curiosity-driven motor primitives and the retreat actor by managing novelty inputs: whenever novelty is too high, the retreat primitive is switched on; if novelty is too low, the next level loop is invoked (Fig. 3).

Formally, the probability of the agent to start returning to base is given by

$$p^l_{\text{retreat}}(r_t) = \psi\left((r_t - (\hat{J}^l_t + \tilde{r}^l))/\tilde{r}^l\right), \tag{12}$$

where r_t is the current reward, $\hat{J}^l_t$ is the current loop's average reward, $\tilde{r}$ is the novelty-transition sensitivity, and $\psi(x) = 1/\left(1 + e^{-x}\right)$ is a sigmoid function. This transition means that as the reward exceeds the current estimation of the average reward, there is a greater probability for a transition to the the retreat primitive. In other words, whenever novelty is significantly greater than the expected (average) novelty, the agent returns to base, where the difference should be greater than $\tilde{r}^l$ in order to reduce the effects of reward-related noise. This transition entails neophobia, i.e. the novelty-averse reaction of the agent, in contrast to the curiosity drive.

Whenever the currently active loop no longer produces novelty, the novelty management unit switches to a higher level loop in order to increase novelty input. The probability for this advance to a higher loop is given by:

$$p^l_{\text{adv}}(\tau^l) = \psi\left((\tau^l - \hat{\tau}^l)/\tilde{\tau}^l\right), \tag{13}$$

where τ^l is the accumulated time in which loop l has been active, $\hat{\tau}^l$ is the advancement threshold and $\tilde{\tau}^l$ is the advancement sensitivity. τ^l is accumulated from the first time-step loop l has become active and is reset whenever it becomes inactive, i.e. whenever there is a transition to another loop or motor primitive. The advance is preceded by a return to base, so that the higher loop is active during the next new entry (see below). This ensures that all loops have a common starting point and learn different features of the environment due to their distinguishing novelty-related aspects and not due to sporadic initial exploration state.

Taking Eqs. (12)–(13) together means that the agent performs the transition to a higher loop when it moved $\hat{\tau}^l$ time steps without receiving a higher-than-expected reward. This can be understood from the fact that if at one point the reward exceeded the estimated average reward, it would have returned to base; hence only if $\hat{\tau}^l$ time-steps it received less than $\hat{J}^l$, it performs the transition to a higher loop. $\hat{\tau}^l$ also gives indication on the minimal timescale of environmental change dynamics, since any change occurring on a shorter timescale will cause the system to be "stuck" in one curiosity loop and not be able to continue to other novel features of the environment.

Loops' timescales Each curiosity loop has its own adaptive timescale, T^l. The timescale influences several processes within and between loops: (i) The average reward temporal window is set such that $1/\xi^l = T^l$, i.e. $\hat{J}_t$ is calculated over a period of T^l time-steps; (ii) The critic and actor learning rates are set such that $\alpha^{(C,A),l} = 0.1/T^l$, i.e. each AC-module has its own learning rate; (iii) Advance time is equal to the loop's timescale, $\hat{\tau}^l = T^l$, i.e. a low-level curiosity loop advances to a higher one after T^l time-steps with no novelty.

The timescale is an adaptive parameter such that it accommodates different agent-environment interactions: an

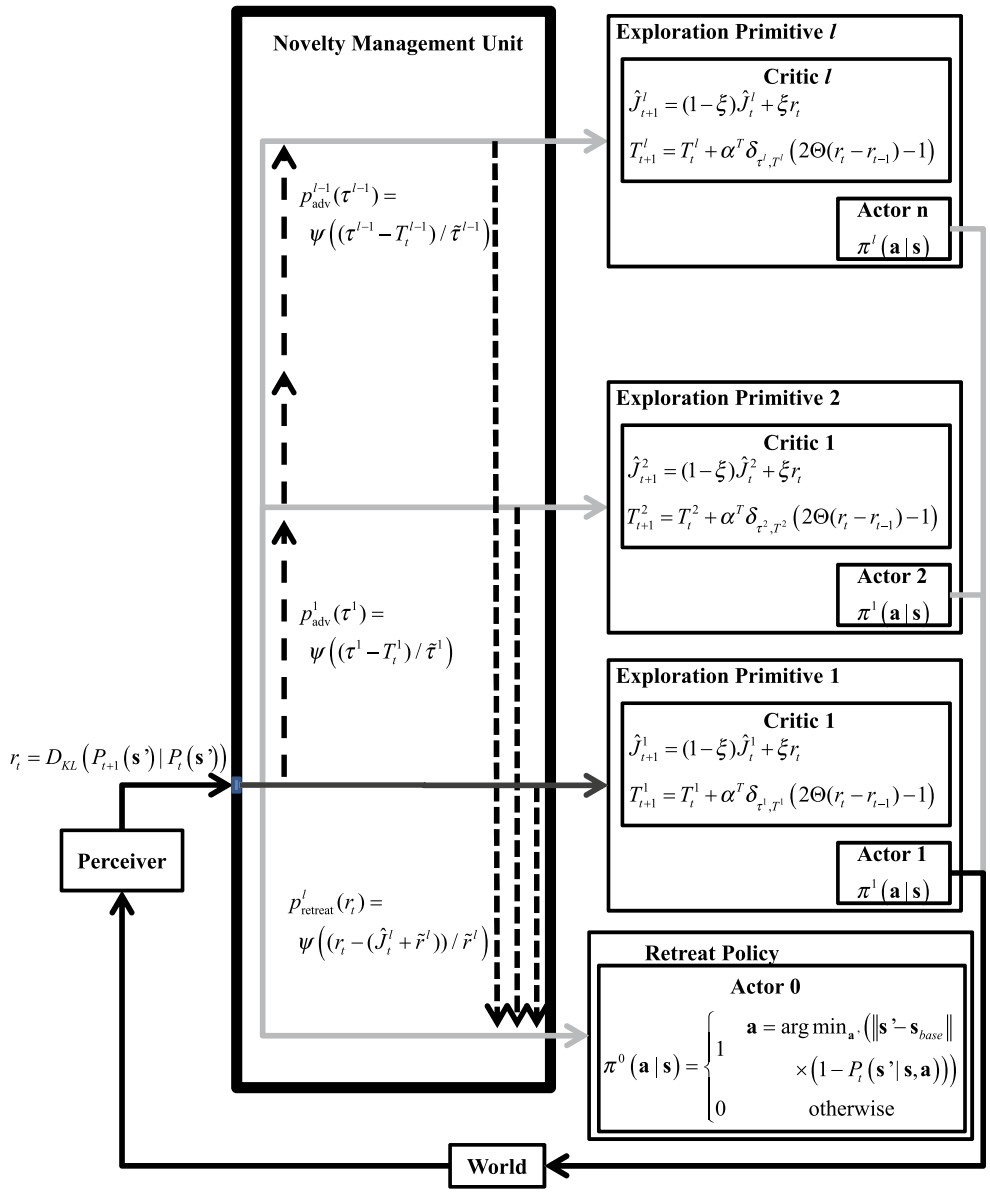

Fig. 3 Hierarchical curiosity loop architecture. A single perceiver calculates the information gain, that goes to the novelty management unit as intrinsic reward. The novelty management unit switches probabilistically between the curiosity loops and the retreat primitive according to this reward: the greater the reward is from the expected (average) reward, which is updated in each exploration primitive's critic, the greater the probability the retreat primitive becomes active (*dotted arrows*); the longer the time no novelty is from the current loop's time-scale, which is updated in each exploration primitive's critic, the greater the probability the next curiosity loop becomes active (*dashed arrows*)

enriched interaction may necessitate more exploration time for certain loops and less for others; conversely, a curiosity loop active in a deprived environment may sometime require a longer duration until it encounters the first novel feature. Thus, the time-scale is updated according to the following dynamics:

$$T^l_{t+1} = \max\left(T^l_t + \alpha^T \delta_{\tau^l, T^l}\left(2\Theta(r_t - r_{t-1}) - 1\right), T^{l-1}_t\right) \tag{14}$$

where α^T is the timescale update rate, δ_{τ^l, T^l} is Kronecker's delta and $\Theta(\cdot)$ is the Heaviside step-function. Eq. (14) reflects the following heuristics: at time $\tau^l = T^l$, i.e. when loop l has been active for a duration that is equal to its timescale, if the current reward is greater than the previous reward, increase the timescale; otherwise decrease it. Furthermore, loop's l timescale cannot be below the previous level's timescale. The logic behind these principles is the following: if the loop has been active for T^l_t time-steps and still receives large rewards, then there is more to discover, and the timescale increases; on the other hand, if it has been active long enough and rewards are diminishing, then the loop should be less active and reduces its own timescale. Moreover, we assume that the hierarchical nature of the curiosity loop demands that the timescales increase with higher loops, as each loop explores a higher dimensionality of the exploration space, thus requiring longer activation times to complete the exploration.

The AC-module learning rates depend on the loop time-scale and are hence also adaptive. The rationale is that

the AC-modules should average over many environments, i.e. the behavior that emerges should be applicable to a wide range of environments. Hence, the change of the AC-module parameters should be slower than encountering a single environment. Each level encounters a single environment on average T^l time-steps; hence fulfilling the requirements for the learning rates amount to $\alpha^{(C,A),l}T^l << 1$. We thus take the learning rates to be $\alpha^{(C,A),l} = 0.1/T^l$, which both fulfills the upper boundary for averaging out environments and is not too small to have too slow convergence.

4.5 Framework parameters

While it appears at first glance that the framework has many parameters, most are linked and adaptive, such that there remain only few free parameters. The parameters of the framework are partitioned to two groups, namely, curiosity loop parameters and novelty management parameters, Table 1. The curiosity loop parameters are: (i) learning and update rates of the actor, critic and timescale, $\alpha^{(C,A,T),l}$, which are set to be equal $\alpha^{(C,A,T),l} = \alpha^l$, where $\alpha^l = 0.1/T^l$ is the adaptive loop-dependent general learning rate; (ii) average-reward update rate, which depends on the timescale, $\xi^l = 1/T^l$.

The management parameters include: (i) novelty reduction parameters, namely, $\tilde{r}^l$ the novelty sensitivity, which determines the amount of stochasticity in the exploration-to-retreat transition; (ii) advancement parameters, namely, $\hat{\tau}^l$, $\tilde{\tau}^l$, the advancement threshold and sensitivity, where the latter determines the stochasticity of the advancement transition and the former is adaptive and given by $\hat{\tau}^l = T^l$.

The two loop-specific parameters, ξ^l and $\hat{\tau}^l$ dictate a timescale for each loop; the former dictates the running average window of the rewards and the latter the no-novelty time before advancement. We have set $\hat{\tau}^l = 1/\xi^l = T^l$ for all loops, such that each loop has a single adaptive timescale, Eq. (14). This means that if by the time the average reward has been calculated, there was no higher-than-expected reward, the next loop should be activated. In other words, advancement occurs only if the rewards are monotonically decreasing for T^l time-steps.

To summarize, all learning rates are adaptive and equal to α^l. Furthermore, sensitivities parameters, $\tilde{r}^l$, $\tilde{\tau}^l$ which control the noise in the transition between primitives are set *a-priori*, for example $\tilde{r}^l$, $\tilde{\tau}^l = 0.01$ represent deterministic transitions. Hence, they are not free parameters that are fitted to the data. All time-scales are adaptive, meaning their converged values do not depend on their initial arbitrary value. Any instantiation of our framework that models a specific environment and modality, e.g. whisking or locomotion systems, may contain model-specific parameters, such as sensory uncertainty and perceiver prior. These can be thoroughly analyzed *a-priori*, thus describing the possible repertoire of emergent behaviors.

Table 1 Framework parameters

Symbol	Parameter	Value	Comment
T^l	Loop timescale		Adaptive
ξ^l	Average-reward update rate	$1/T^l$	
$\hat{\tau}^l$	Advancement threshold	T^l	
α^l	Loop learning rate	$0.1/T^l$	
α^C	Critic learning rate	α^l	
α^A	Actor learning rate	α^l	
α^T	Time-scale update rate	α^l	
$\tilde{r}^l$	Novelty sensitivity		Transition noise
$\tilde{\tau}^l$	Advancement sensitivity		Transition noise

4.6 Exploration sessions

Exploration sessions start from the base state and are divided to entries and excursions. A new excursion is characterized by reseting the perceiver and a randomization of the environment. It signifies exploration of a new environment, which can be of variable configuration, and thus requires a reseting of the perceiver, since it can no longer comply with the different size, shape or other environmental parameters. A reminder: the perceiver is not invariant, i.e. it depends on the entire state-space and thus must be reset when the state-space changes. The first active actor at each new excursion is the first level's actor, π^1.

A new entry is characterized by the same environment as the previous entry, i.e. the environmental parameters do not change. Hence, the perceiver is not reset, but rather continuously updated during exploration. The active actor at the beginning of a new entry is set to be the last actor prior to the retreat, $\pi^{l_{\text{last}}}$, in order to maintain continuation of order-of-loops execution.

A new excursion, i.e. exploration of a new environment, begins only after the last loop has passed its advancement threshold, i.e. whenever all the loops have not acquired new information about the current environment. The introduction of new environments is not part of the internal curiosity model, but rather represents the encounter of the agent with a changing world. It is an external imposition, and is set to make exploration policies, i.e. motor primitives, converge to more robust behaviors that generalize over many environments.

An example exploratory excursion The first actor in an excursion is always the first-level actor, $\pi^1(a|s)$. It explores

the environment and switches to the retreat primitive whenever it receives more-than-expected reward. After reaching the base-state, a new entry begins with the same first loop's actor. Only after it has learned the most rewarding feature of the environment, and has received no "surprising" rewards, it advances to the next loop. The agent returns to the base-state and starts a new entry with the higher loop's actor. The new active loop explores with the same perceiver as the first loop (the agent has only one perceiver), hence previously encountered features are no longer rewarding. The second actor is active until it encounters a novel feature, which instigates a more-than-expected reward. This is followed by a transition to the retreat primitive, until the agent reaches the base-state. At this point, a new entry begins, with the same perceiver and the active actor is the second actor (the one that was active prior to the retreat). It alternates between exploration and retreat until it does not encounter any rewarding or novel signals. At this point, it advances to the next loop. When the last loop has encountered no rewarding signal, the agent is ready to seek a new environment.

5 Implementation in the whisking system

We implemented the framework to model the rodents' whisker system, which is used as an active sensing modality to explore nearby surrounding (Szwed et al. 2003; Kleinfeld et al. 2006). We first introduce the neurophysiological basis, model assumptions and framework implementation. It is followed by simulation results and concluded with relation of results to observed whisking behavior and novel predictions.

5.1 Whisking system model

Neurophysiological basis Whiskers are actively controlled by two muscle groups, namely, intrinsic muscles (Berg and Kleinfeld 2003; Hill et al. 2008; Simony et al. 2010) and extrinsic muscles (Berg and Kleinfeld 2003). The former connect two adjacent whiskers of the same row (Simony et al. 2010) and the latter are connected to the mystacial pad (Berg and Kleinfeld 2003). It was reported that protraction is mainly performed by the intrinsic muscles and one of the external muscles, while retraction is either passive or affected by extrinsic muscles during palpation and exploratory whisking, respectively (Hill et al. 2008; Berg and Kleinfeld 2003).

The information from the whisker follicle is conveyed via the infra-orbital nerve to the trigeminal ganglion (TG)(Szwed et al. 2003; Szwed et al. 2006; Leiser and Moxon 2007). There are three types of sensory neurons in the TG, namely, whisking, touch and whisking-touch which respond mainly during whisking, contact with objects or both, respectively. The touch neurons in TG can be subdivided to contact, pressure and detach cells, which respond to the initial contact, the prolong pressure during contact and the detachment from the object, respectively (Szwed et al. 2003).

Assumptions We model the whisker as a one-dimensional agent characterized by the normalized azimuth angle $\theta_t = 0, \Delta\theta, \ldots, 1$ (Knutsen et al. 2008), where $\theta_t = 0$ and $\theta = 1$ denote full retraction and protraction, respectively. The angle is discretized to $N_\theta = 1/\Delta_\theta + 1$ angles. In each new excursion, the whisker field size is chosen randomly, $N_\theta \in [7, 13]$, signifying a variable environment and emphasizing the fact that the actors and critics are invariant to whisker field size. Furthermore, the whisker is treated as a rigid body, i.e. it cannot bend or "penetrate" objects.

Dynamics The whisker dynamics is governed by the discretized change of the angle, driven by retraction and protraction commands, $a_t = \{-1, 1\}$, respectively. In our simplified model, the dynamics is linear and have the following form: $\theta_{t+1} = \theta_t + a_t \Delta\theta$, where θ_t is always bounded by 0 and 1.

Upon contact with an object, since the whisker is treated as a rigid body, $\theta_t \leq \theta_o$, where θ_o is the position of the object in azimuth angle. Hence, even when protracting, the whisker cannot "penetrate" the angle of the object, which results in angle absorption (Szwed et al. 2003), i.e. the difference between the angle if there had not been an object and the true angle. We model the touch information by a binary variable, $\gamma_t \in \{0, 1\}$ which equals one as long as there is contact with an object, i.e. whenever protracting against the object; zero otherwise.

State/action spaces In the implemented whisker model, the sensory information is composed of whisker angle, θ_t, as conveyed by whisking cells (Szwed et al. 2003; Szwed et al. 2006; Leiser and Moxon 2007), and contact information, γ_t, as conveyed by touch cells (Szwed et al. 2003). The actions are modeled as either protraction or retraction, a_t.

Since the whisker angle is bounded and discretized, it is considered as the *variable* part of the state-space, $\check{s} = \{\theta_t\}$ and has a well defined metric. Thus, in order for the motor primitives to be transferable, they cannot depend on θ_t. The touch information, on the other hand, is binary, making it the invariant part of the state-space. The critics and actors can only depend on those.

We also consider the previous touch information, γ_{t-1}, where initial touch information is taken to be zero, i.e.

$\gamma_{-1} = 0$ meaning there is no contact with an object when the whisker is fully retracted and starts moving. By inclusion of the previous touch information, all three types of touch cells (Szwed et al. 2003) can be represented, where we introduce the notation of the sensory neurons, $\gamma_t^N = \{C, P, D, W\}$ for contact, pressure, detach and whisking cells, respectively, Table 2

5.2 Whisker curiosity loop

Whisker perceiver The perceiver of the curiosity loop attempts to predict the next sensory state of the agent. In the case of the whisker dynamics, the state-space describes the angle and contact information, $\mathbf{s}_t = \{\theta_t, \gamma_t\}$. The perceiver is thus defined as $L(\theta_{t+1}, \gamma_{t+1}|\theta_t, \gamma_t, a_t)$, i.e. attempting to predict the next activity of the whisking and touch cells.

In many applications of Bayes updates (Miyazaki et al. 2006), there is a spatial (or other) relation between states, such that the greater the "distance" between the observed and true states, the smaller the probability $q(\mathbf{o}|\mathbf{k})$. Usually, this acquires a Gaussian form: $q(\mathbf{o}|\mathbf{k}) = Z^{-1}\exp^{-||\mathbf{o}-\mathbf{k}||/\sigma^2}$, where Z is a normalization factor, $||\cdot||$ is the distance between the states and σ is the width of the Gaussian, representing the noise. This representation can be used only for that part of the state space that has a geometric relation, e.g. whisker angle or arena location. However, such a metric cannot be a-priori assumed for all the state space, e.g. contact information has no a-priori metric. Moreover, defining a single metric for the entire state-space is problematic, at best.

Hence, we do not assume any such metric, but rather define a dichotomic relation that singles out the correct state $\mathbf{o} = \mathbf{k}$ and equates all the others. We then set, for a discrete state-space:

$$q(\mathbf{o}|\mathbf{k}) = \begin{cases} \frac{1}{1+\sigma} & \mathbf{o} = \mathbf{k} \\ \frac{\sigma}{1+\sigma} & \forall \mathbf{o} \neq \mathbf{k} \end{cases} \tag{15}$$

where σ represents the noise, $\hat{\sigma} = \sigma/(N_s - 1)$ is the normalized noise and N_s is the number of discrete possible observed states. $\hat{\sigma}$ conveys how much the observed sensory information is unreliable: a very low $\hat{\sigma} \ll 1$ means that given an actual state $\mathbf{k}$, it is highly likely that it will be observed correctly $\mathbf{o} = \mathbf{k}$, while other states have very low probability to be observed; if $\hat{\sigma} = 1$, all states have the same probability to be observed, irrespective of the true state. In this formulation there is no increased probability for "close" states since there is no *a-priori* metric that maps them.

Table 2 Representations of whisking and contact cell types

Cell type	Notation	γ_{t-1}	γ_t	Comments
Whisking	$\gamma_t^N = W$	0	0	no contact at all
Contact	$\gamma_t^N = C$	0	1	no contact followed by contact
Pressure	$\gamma_t^N = P$	1	1	maintained contact
Detach	$\gamma_t^N = D$	1	0	contact followed by no contact

The perceiver Bayes update then acquires the simple form:

$$P_{t+1}(\mathbf{s}'|\mathbf{o}) = \begin{cases} P_t(\mathbf{s}')\frac{1}{\hat{\sigma}+P_t(\mathbf{o})(1-\hat{\sigma})} & \mathbf{s}' = \mathbf{o} \\ P_t(\mathbf{s}')\frac{\hat{\sigma}}{\hat{\sigma}+P_t(\mathbf{o})(1-\hat{\sigma})} & \forall \mathbf{s}' \neq \mathbf{o} \end{cases} \tag{16}$$

In the whisking system model, at each time step t, the agent updates its probability map L by using its acquired current and previous time-step information, namely, $\theta_t, \theta_{t-1}, \gamma_t, \gamma_{t-1}$ and a_t. The uncertainty of the update is given by σ, Eq. (16). It determines what is the probability that the true angle and contact are perceived by the correct whisking and contact cells; a low σ means that the correct cells have higher probability of being activated than incorrect cells (those that represent different angles and contact information). It is important to note that while the angle has a well-defined distance metric, contact information does not. Thus, it is difficult to ascertain a whole state-space distance, e.g. a distance between two states $||\mathbf{s}_1 - \mathbf{s}_2||$. For this reason, the noise function Eq. (15) was used instead of a more common Gaussian-based noise model.

Whisker critics and actors The invariant critics and actors depend only on the extended invariant parts of the state-space, $\tilde{s} = \{\gamma_t^N\}$, i.e. they depend only on which cell is active, the whisking or touch cells. The critics, $\hat{V}^\pi(\gamma_t^N; \nu_t)$, predict the value given the current cell activation, but not the value's dependence on the current whisker angle (since it is not invariant). The actor is also angle independent, wherein the probability to protract or retract depends only on cell activation identity, i.e. whisking/contact/detach/pressure.

We summarize the tunable parameters of the critic and actor: the critic, $\hat{V}^\pi(\gamma_t^N; \nu_t)$ is represented by $||\nu_t|| = 4$ values and the actor, $\pi(a_{t+1}|\gamma_t^N; \lambda_t)$ is represented by $||\lambda_t|| = 4$ protraction probabilities.

Furthermore, we are using relatively simple representations for the critic and actor: $V(\mathbf{s}; \nu) = \nu_{\mathbf{s}}$, $\pi(\mathbf{a}|\mathbf{s}\ \lambda) = \lambda_{\mathbf{s}}^2/\sum_{\mathbf{k}}\lambda_{\mathbf{k}}^2$.

5.3 Whisker novelty management

We implemented a two-loop model, $N_H = 2$, wherein each loop contains a critic, $V^{1,2}(\tilde{s})$ and an actor, $\pi^{1,2}(a|\tilde{s})$. We assume that the first loop is insensitive to the contact

information, $\gamma_t^N = \{C, P, D\}$, and only depends on whisking cell activation $\gamma_t^N = \{W\}$, hence the critic has a single value and the actor has only a single protraction/retraction probability. This assumption guarantees that the first loop learns only whisker dynamics and not whisker-object interaction (Gordon and Ahissar 2012). The second loop depends on the entire invariant state-space, i.e. whisking and touch cell activation. In order to represent these assumptions numerically, we have set that entries starting with the first-level loop contain no objects in the whisker field, whereas entries that start with the second loop may have objects. In each new excursion there is a p_{obj} chance of objects being presented in a random position along the whisker-field.

Retreat policy In the aforementioned simplified whisker model, the base-state is set to be the fully protracted whisker, i.e. $\check{\mathbf{s}}_{\text{base}} = \{\theta = 0\}$. Due to the binary nature of the actions, i.e. either protraction or retraction, the retreat actor also assumes a simple form, namely, it performs retraction until reaching the base state.

Model parameters For the initial prior before the first run of the algorithm P_0, we take a uniform distribution for the perceiver, i.e. there is equal probability for every angle and contact information, irrespective of the current angle or action performed.

5.4 Results

We start with a simple example run that illustrates the complex dynamics of the hierarchical loops and novelty management. It is followed by an analysis of the convergence process, in which the motor primitives are learned. We conclude with the description of the converged behavior of the agent.

Example dynamics We present an excerpt of the full dynamics, after more than 6000 time steps, beginning with a new excursion. The whisker initially starts in the base state of a fully retracted whisker, Fig. 4. The learner has no information about the whisker dynamics, i.e. each new action-state transition is novel. The initial actor is the first level actor (Fig. 4a, gray), which has an increased protraction probability (not shown), i.e. given the current state, the actor probabilistically determines the next action, here protraction. The previous state, current action performed and current state provide information for the perceiver, which generates the Bayesian update, Eq. (16). Reward is then calculated as the information gain of the update, Eq. (10) (Fig. 4b, black), which is high whenever the agent encounters a new state-action transition. This reward affects two processes: first, the reward is passed to the critic, which compares the difference between its prediction of the reward of the previous state and the current state to the difference between the average reward and the received reward and calculates the TD-error, Eq. (1). Based on the TD-error the critic values and the actor probabilities of the current state are updated, thus completing a single curiosity-loop cycle.

Once the loop receives a reward higher than the average reward (Fig. 4b, dark/light gray), the novelty management unit switches to the retreat actor (Fig. 4a, black), retracts the whisker back to the base-state and the whisker begins a new entry. The first loop's actor continues until another novel

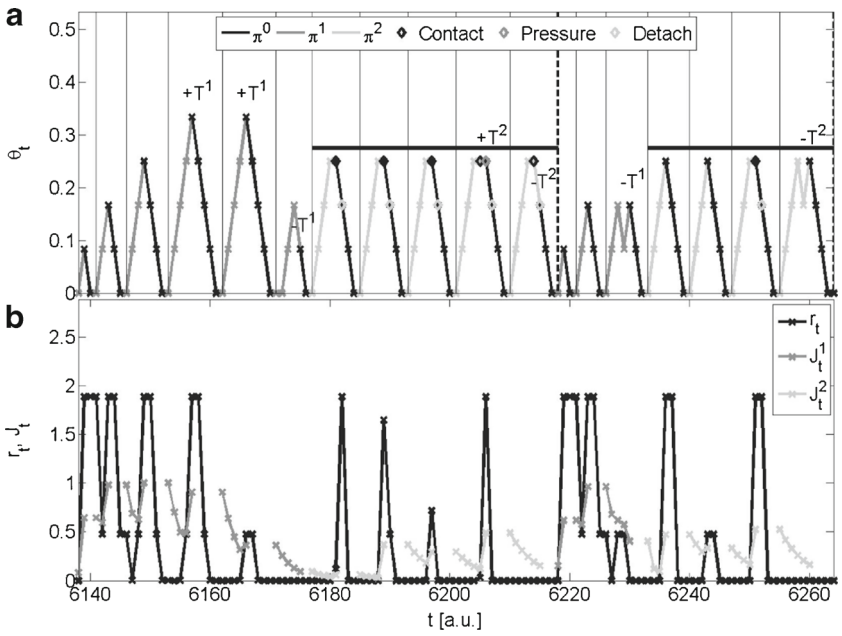

Fig. 4 An example run of whisker dynamics. **a** Whisker angle as a function of time, color coded for active actors: first loop actor coded as *gray*, second loop actor coded as *light gray*, retreat actor coded as *black*. Touch cell activation are shown in *diamonds*: *black* for contact, *gray* for pressure and *light gray* for detach cells. *Vertical solid lines* indicate a new entry; *vertical dashed lines* indicate a new excursion. *Horizontal lines* indicate position of object. $\pm T^l$ indicates the time of increase/decrease in timescale of loop l. **b** Reward r_t (*black*) and first (*gray*) and second (*light gray*) level average rewards $\hat{J}_t^{l,2}$ as a function of time. Parameters: $\sigma = 0.5$, $p_{\text{obj}} = 1.0$

action-state transition is reached, whereupon the retreat primitive is activated, and so on.

The loop's time-scale is updated after T^l time-steps have passed, whereupon it increases if the reward increased in the last step or decreases otherwise, Eq. (14). Only after T_t^1 time-steps have passed without any novelty while the first loop's actor behaves, the novelty management unit switches to the second level actor (Fig. 4a, light gray). When the second actor is active, objects may be present in the whisker field (Fig. 4a, horizontal lines). The first time the whisker touches an object, the contact cells become active, which is a novel event that causes full retraction back to the base-state (Fig. 4(at), gray diamond). The retraction itself, although occurring with the retreat actor, activates the detach cells (Fig. 4(a), black diamond) and teaches the perceiver about its presence. Contacting the same object again is slowly becoming less novel, which ends the second-loop activation, followed by a new excursion.

Loop convergence We now analyze the convergence of the loops as a function of time, i.e. as the whisker makes more and more entries, while exploring its own whisker dynamics and its object-filled environment. Several parameters change with interaction with the environment, the most notable are those that affect the behavior, namely, the actor's protraction probabilities (Fig. 5a), loops' timescales and average reward (Fig. 5b). The actor's probabilities converge in a cascade of loops, i.e. the first loop's free-air behavior converges first to full protraction, followed by the second loop's behavior. The latter can be characterized by the following sequence: if there is no contact, protract; upon initial contact, continue protraction instigating the activation of the pressure cells; upon pressure, retract instigating the activation of the detach cells; upon detach either protract or retract with half-half probability, completing the cycle of object palpation.

The convergence of the loops' timescale (Fig. 5b) shows that both loops converge to the same value, which is the span of the whole whisker field. The first loop's timescale determines the self-motion exploration, i.e. scanning the protraction-retraction-angle correlation. The second loop's timescale determines the maximal position of objects, which can appear within the entire whisker field. Furthermore, the loops' average reward also converge to a similar value, indicating that the flow of novelty, measured by bits-per-timestep, are similar across loops.

Exploration efficiency The efficiency of exploratory behavior can be measured along two axis, namely, the time of exploration and the perception of the explored environment at the end of exploration. The efficiency of the converged policies, in terms of time and accuracy of perception, is compared to a random behavior (i.e., without prior learning in previous environments), both combined with the novelty-management principle (Figs. 6). The results show that converged loop 1 is slower than random action in perceiving the whisker self-dynamics (Fig. 6a, left), but more accurate (Fig. 6b, left). This is due to the fact that the transition between exploration primitives is governed by the heuristics that after T^l time-steps with low novelty, a higher loop is activated. Hence, in the random behavior there is a greater chance of performing a "random walk" in whisker-space that will generate low-novelty; in other words the random-moving whisker does not explore the entire whisker space before it perceives there is no more novelty. On the other hand, the converged self-motion exploration motor primitives covers the entire whisker-space before moving to a higher level, thus taking it

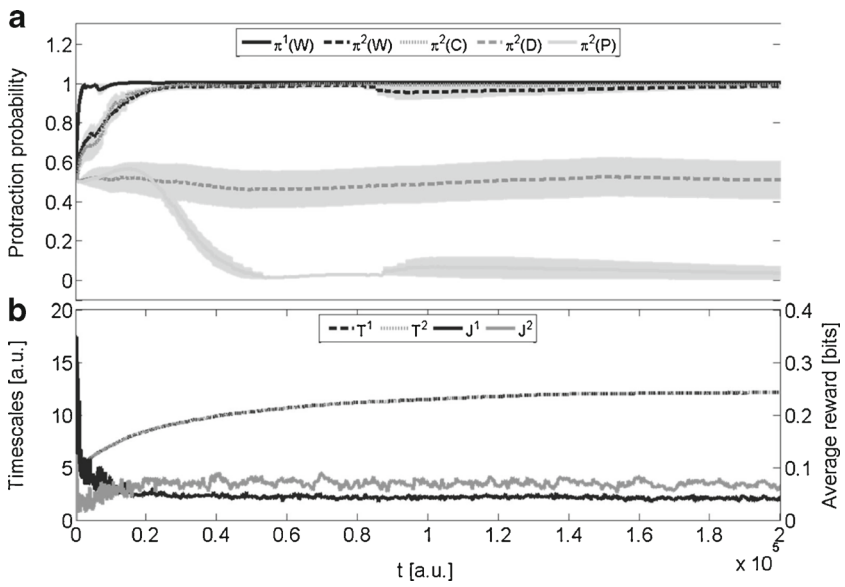

Fig. 5 Convergence process, averaged over 10 runs. **a** Mean protraction probabilities for first and second actors as a function of time. First loop's actor does not depend on contact information, hence is represented by $\pi^1(W)$; second loop's actor protraction probability $\pi^2(W/C/D/P)$ depends on whisking, contact, detach and pressure, respectively. Shaded areas represent standard error. **b** First and second loops' mean timescales and average reward as a function of time. Same parameters as Fig. 4

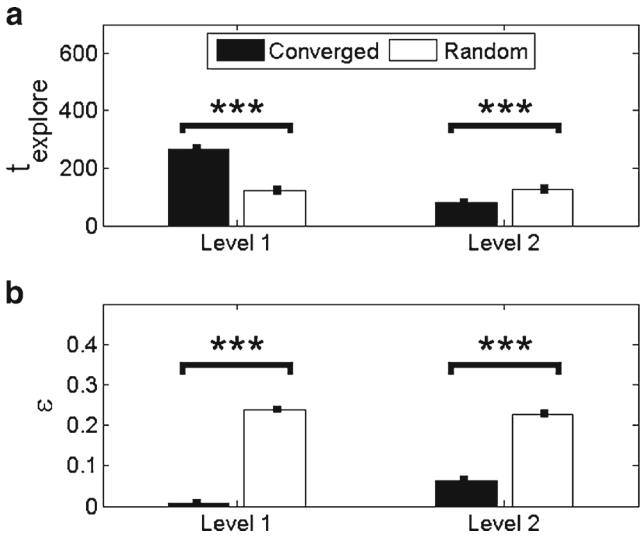

Fig. 6 Comparison of exploration efficiency between the converged actors and random actors, both with novelty management. **a** Exploration times of loops 1 and 2 for the converged and random actors measured as the number of time-steps before releasing control to the next loop. **b** Perceiver error at the end of exploration, measure as the mean-square-error between the perceiver and the true whisker map. Same parameters as Fig. 4

more time, but achieving a higher-accuracy perception of self-motion.

Converged loop 2 perceives objects faster (Fig. 6a, right) and more accurately than its random counterpart (Fig. 6b, right). This occurs since the objects are localized in whisker-space and random movements have low probability of encountering and palpating their borders. The object-localization motor primitive, on the other hand, seeks out objects in the optimal manner by protracting until touch, and then palpating the object borders in an optimal manner, in the sense that perception of all touch-cell activations is explored in the least amount of time. Hence, the converged motor primitive is much more efficient than randomly moving the whisker.

Whisking behavior We first describe the model's explicit results, shown in Fig. 7. While the figure represents a single example, the convergent exploration motor primitives were practically identical for 10 separate runs (Fig. 5a), and for all parameters space ($p_{\text{obj}} \in (0, 1]$ and $\sigma \in (0, 1)$, not shown) hence exhibiting the same behavior as the one presented here. In the next section, we qualitatively compare these behaviors to reported behaviors of rodents' whisking and suggest a novelty-driven explanation, Table 3.

To fully represent a natural scenario, the example in Fig. 7 starts with unknown whisker dynamics and continues with an environment that may change after the end of the excursion, i.e. a new object may appear in a random position in the whisker field. The changed environment does not instigate re-learning of the whisker dynamics via the first loop, but rather the second (and last) loop continues to be active and explore the reappearing objects.

In the beginning of the event (Fig. 7(B1)), the first motor primitive is active, resulting in protraction. However, since it has no information on the whisker dynamics, whenever it encounters a new state-transition of the system, i.e. a new whisker angle is reached, novelty rises and with it a large reward. In this example, the sensory uncertainty is low such that one encounter with the new state suffices to update the learner in such a way that encountering the same state-transition the second time does not instigate a large enough reward. Hence, the recurrent activation of the first motor primitive results in a gradual sequential exploration of the whisker field. It ends when the entire whisker field has been explored and one sweep produces decreasingly smaller rewards. This results in a retreat and activation of the second motor primitive.

In this example, initially there are no objects in the whisker field (Fig. 7(B2)). Hence, the second motor primitive protracts until it too sweeps the entire field, with no encountered novelty. This results in the end of an excursion and a retreat. However, as mentioned before, here the whisker starts again with the second motor primitive and does not re-learn the whisker dynamics with the first motor primitive. While the free-air component of the two primitives is identical, due to the already-acquired information about the whisker dynamics, there is no larger-than-expected novelty and hence no return with full retraction. The whisker continues to explore, alternating between full protraction and full retraction, until it encounters an object, i.e. until the contact cells are activated.

When the whisker first encounters an object, there is a rise in novelty and hence reward is greater than expected, resulting in immediate retraction due to the retreat primitive (Fig. 7(B3)). Subsequent contacts continue to explore the object by the full activation of the second motor primitive (Fig. 7(B4)): contact is followed by protraction, resulting in pressure-cell activation; pressure is followed by retraction, resulting in detach-cell activation; detach is followed by either retraction, resulting in whisking-cell activation, or by protraction, resulting in re-activation of the pressure cells; whisking-cells induce protraction causing another palpation cycle. This touch-induced behavior ends with full retraction either when there is large novelty due to encountering a new touch-cell activation, or if enough time has passed with no higher-than-expected rewards. In the latter case, the whisker starts a new entry with the second primitive, seeking a new object.

If the whisker palpated an object that was then removed, it considers the non-activation of the contact cell at the object's previous location as novelty, since it expected the

Table 3 Whisking behaviors and their model counterparts

Behavior	Model	Principle	Comments
B1. Whisker twitching (Semba et al. 1980; Nicolelis et al. 1995) (Fanselow et al. 2001)	Alternation between loop 1and Retreat	Novelty-managed exploration of whisker dynamics	Immobility reduces environmental variation
B2. Exploratory whisking (Gao et al. 2001; Berg and Kleinfeld 2003) (Towal and Hartmann 2008; Knutsen et al. 2008)	Alternation between loop 2 and Retreat	Novelty-poor contact-seeking entries	Independent of sensory information
B3. Rapid cessation of protraction (Mitchinson et al. 2007; Grant et al. 2009)	Retreat retraction upon initial contact	Novelty-aversive of reaction to first touch-cell activation	
B4. Touch–induced pumps (Deutsch et al. 2012)	Loop 2 exploration primitive	Novelty-seeking exploration of touch–cell activations	Less pumps during initial contacts, more pumps during earlier whisking cycle

object to be there (Fig. 7(B5)). Hence, this novelty induces a retreat activation resulting in a "phantom touch"-induced behavior. The surprise lasts another sweep and the next time it arrives to the same location, it no longer expects the object there and continues to explore the whisker field.

5.5 Behavioral comparison

The initial behavior of incremental exploration of the whisker field, intermittent with full retraction is reminiscent of "whisker-twitching" (Semba et al. 1980; Nicolelis et al. 1995; Fanselow et al. 2001), which is described as high-frequency small amplitude movement of the whisker during attentive immobility, i.e. when the rat does not move its head or torso but is awake. The model's behavior is the result of the first motor primitive activation which facilitates learning the internal whisker dynamics. In relation to a freely moving animal, one may suggest that it is beneficial to do so while not encountering objects in the environment, e.g. by restricting the mobility of the torso and head.

The second behavior is exploratory whisking (Gao et al. 2001; Berg and Kleinfeld 2003; Towal and Hartmann 2008;

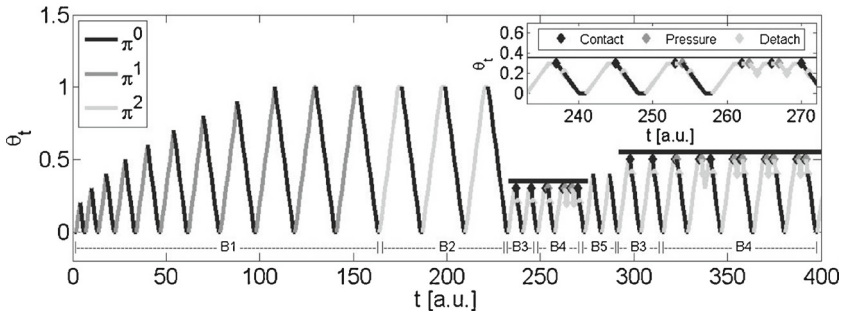

Fig. 7 Behavior of the entire converged model; whisker angle (θ_t) is depicted as a function of time, where color denotes the active actor (same color code as Fig. 4). *Black horizontal lines* denote the angular position of an object. B1, actor 1 protracts the whisker and the retreat primitive retracts the whisker whenever a new angle is reached. B2, initially there are no objects in the whisker field and it protracts, whereupon experiencing no novelty, the novelty management unit switches to the retreat policy (retraction). When objects are present, the initial contact is novel and immediately followed by retreat (B3), whereas the following contacts slowly exhibit the full dynamics of the converged actor 2: protract upon contact (*black diamond*), retract upon pressure (*gray diamond*) and either protract ($t = 264$) or retract ($t = 337$) upon detach (*light-gray diamond*) mechanoreceptor activation (B4). B5, when an object is removed from the whisker field, retreat follows high novelty due to false prediction of its location. Inset: Zoom-in on the initial contact with an object. Same parameters as Fig. 4

Knutsen et al. 2008), wherein the second motor primitive is active when there are no objects. The rhythmic whisking behavior in the model is the result of intermittent protraction due to the second motor primitive when there are no touch-cell activations and the retreat full retraction. The frequency of this periodic behavior is mainly controlled by the time-scale variable of the second loop, T_2, with increased frequency represented by a shorter time-scale. However, it is important to note that in the model, periodic exploratory whisking is a learned behavior, as exemplified by the convergence of the second loop's actor.

When an object is first encountered, the model suggests that the first reaction is full retraction, due to the high novelty of the whisker-object interaction. This behavior was observed and dubbed "rapid cessation of protraction" (Mitchinson et al. 2007; Grant et al. 2009) and occurred, as in the model, only during the initial contact with the object. While the full reported behavior also describes the contralateral whisker increased protraction, our model currently implements a single whisker and cannot account for this result (see below).

During following contacts with the object, the model shows that during the activation of the second motor primitive, the object is palpated by discrete events of contact, prolonged contact, detachment and another induced contact. This behavior was also observed and called "touch-induced pumps" (TIP) in Ref. (Deutsch et al. 2012), which also performed a detailed analysis of the occurrence of the pumps. The model predicts that TIPs should occur less during the initial contacts with the object, due to the high novelty content of first contacts and the following novelty reduction via the retreat primitive. The model further predicts that due to the adaptive time-scale of the second curiosity loop, i.e. how long with no novelty can a motor primitive be active before retreating, objects encountered during latter stages of whisking should produce less TIPs. This happens, according to the model, since arriving to a more distant object entails more time with low novelty and only a few TIPs can occur before the temporal threshold $T^{(2)}$ is reached. Mechanistically, the model suggests that the retraction during a TIP, which happens due to the second-level motor primitive, is different than that of exploratory whisking, which happens due to the retreat primitive. For example, retraction speeds may be different. All of these predictions were observed and analyzed in Ref. (Deutsch et al. 2012).

To conclude the behavioral section, although our whisker model is a simplified version of a much richer whisker dynamics (Hill et al. 2008; Simony et al. 2010), the emerged patterns from the hierarchical curiosity loops architecture qualitatively convey many reported behaviors. Nevertheless, some known whisker motion patterns and details are not reproduced by the current model, but may be reproduced by simple extensions.

Predictions A behavior that emerges from the model is that when objects are rapidly removed from the whisker field, the whisker behaves as if the object was there, i.e. it retracts upon "mis-contact" or "phantom-touch". While one may interpret it as "predictive whisking", i.e. the rat prepares for the object and thus whisks to that position, our model interprets it differently. The novelty-management principle results in novelty when after the whisker palpated the object several times, thus confirming its position, it is suddenly not there. The "mis-contact" produces novelty, higher than expected reward, thus resulting in retreat behavior. This interpretation can be verified by manipulating the time in which the object disappears: if it disappears after the first contact, the probability for a phantom-touch is much reduced, since according to the model, the rat still has high uncertainty regarding the object position. However, the phantom-touch probability should increase with increased number of palpations. Furthermore, introducing perturbations that increase the whisker noise, or sensory uncertainty, should also drastically affect the appearance of phantom-touches.

The 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 whisk periodically in free-air or 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. 2012).

Furthermore, the converged behaviors are strongly dependent on the experience of the pup, thus changing pups' rearing 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 free-air whisking 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. Moreover, 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.

Whisker model extensions The whisker model we have implemented can be extended in several ways to account for more diverse whisker dynamics. Currently, the model's discretization of the angle and control and the simplistic

dynamics implemented are limited. Extending the model to include more realistic dynamics (Hill et al. 2008; Simony et al. 2010), in which motor neurons' activation results in whisker motion, requires higher-dimension actors and critics. This extension may reproduce a more exact temporal dynamics of whisker movements, e.g. velocity profile, that is analyzed in several studies (Towal and Hartmann 2008; Grant et al. 2009).

The hierarchical curiosity loop architecture has focused on a single whisker, whereas several studies have reported multiple whisker dynamics (Towal and Hartmann 2006; Grant et al. 2009). The model can be extended to include multiple whiskers in a straightforward way by increasing the dimensionality of the state- and action-spaces, e.g. θ_t^i, a_t^i, where $i = 1, \ldots, N_w$, N_w is the number of whiskers. Furthermore, the environment can be made richer and include complex objects, such that the exploration perceiver will eventually result in shape recognition. The emergent motor primitives are hypothesized to include coordinated whisking and "look forward" behavior of the contra-lateral side of touch (Grant et al. 2009).

6 Relation to previous models

The framework builds upon and extends the model presented in (Gordon and Ahissar 2012). The model introduced the basic curiosity loop, which is an intrinsic reward actor-critic framework. However, (Gordon and Ahissar 2012) implemented artificial neural networks, whereas the current implementation is based on the more rigorous Bayesian framework. Furthermore, the intrinsic reward in (Gordon and Ahissar 2012) was taken to be the prediction error of the neural network, whereas we use here the information gain, which is more robust to noise and has a much deeper theoretical foundation and can be actually measured (in bits).

In this contribution, we concentrated on one perceiver per modality, which accommodates several curiosity loops, each converging to exploration of different features of the same perceived environment. In (Gordon and Ahissar 2012), on the other hand, multiple perceivers were introduced, there referred to as forward and inverse models, each accommodating a single curiosity loop. This difference results in emergence of different exploration motor primitives as shown in (Gordon and Ahissar 2012). Furthermore, the proposed framework does not assume which features are explored first and which later on in the hierarchy. The sequence emerges from the interaction with the environment. In (Gordon and Ahissar 2012), on the other hand, the structure of the perceivers higher in the hierarchy was manually constructed.

Finally, both (Gordon and Ahissar 2012) and this contribution implement the model on the whisker system. However, the framework presented here is based on more rigorous mathematical foundations and introduces the notion of novelty management. The repertoire of behaviors explained by the framework is also more extensive and includes whisker twitching, rapid cessation of protraction and touch-induced pumps. Furthermore, novel behavioral and neurophysiological predictions arise from the current framework that were lacking in the (Gordon and Ahissar 2012) model.

Furthermore, several previous models have addressed similar issues and behaviors as those presented here. They are summarized in Table 4, analyzed according to several axis: Emergent vs. pre-programmed; convergent vs. ever-changing behavior; novelty management; hierarchical architecture; external vs. intrinsic motivation and quantitative vs. descriptive.

6.1 Emergent versus pre-programmed or random behavior

Exploratory behavior in specific scenarios can usually be easily programed using a dedicated algorithm, e.g. (Tchernichovski et al. 1998; Tchernichovski and Benjamini 1998; Harish and Golomb 2010). However, designing a

Table 4 Model comparison. Headers indicate: Emg. emergent; Cnv. convergent; Hrc. Cntl. hierarchical control; Mot. motivation; Qnt. quantitative. In-table acronyms: Att./Rep. attraction/repulsion; App./Avoid. approach/avoidance

Model	Emg.	Cnv.	Novelty	Hrc. Cntl.	Mot.	Qnt.
Our framework	V	V	Control	V	Int.	V
(Tchernichovski et al. 1998; Tchernichovski and Benjamini 1998)			Att./Rep.			V
(Harish and Golomb 2010)						V
(Oudeyer et al. 2007; Der and Martius 2012; Little and Sommer 2013)	V		Maximize		Int.	V
(Tishby and Polani 2011)	V		Maximize		Int.+Ext.	V
(Moldovan and Abbeel 2012)			Reduce			V
(Pape et al. 2012)	V		Maximize	V	Int.	V
(Barto et al. 2004)	V			V		V
(Elliot 2006)			App./Avoid.	V	Int.	

single pre-programmed non-adaptive behavior to account for the entire repertoire of behaviors is daunting. Furthermore, it is evident that animals do not behave in a random manner, but rather perform goal-directed actions. The scenarios we address, however, do not have an inherent trivial goal, such as reaching a target or solving a maze and animals do not get an external reward for a specific behavior. We consider scenarios in which they freely behave in a natural intrinsically-motivated manner (Harlow 1950; Schmidhuber 1990; Oudeyer et al. 2007; Singh et al. 2010; Baldassarre 2011). Hence, the formal framework we propose is ubiquitous and can be implemented in any scenario wherein an animal encounters a novel environment, with no external reward. The behavior itself is not defined by the formal framework, but rather emerges from the interaction of the animal with the environment, where the former attempts to learn to predict this interaction.

The interaction between animal and environment is incorporated via the curiosity loop, wherein the perceiver-updates produce the information gain, translated to intrinsic reward that drives the actor-critic algorithm. The actors are thus learned and modified according to novelty, and not according to a specific state or external rewards. We suggest that these behaviors converge to specific motor primitives of exploration, and are not continuously changing, opposed to other suggested models (Der and Martius 2012). Hence, the behavior as a whole converges to a specific, structured yet complex behavior.

6.2 Convergent versus ever-changing policies

There are several relevant time scales that emphasize one of the main differences between the proposed framework and others (Oudeyer et al. 2007; Tishby and Polani 2011; Der and Martius 2012; Pape et al. 2012; Little and Sommer 2013). Here the perceiver is updated during a *perceptual time scale*, i.e. during exploration of a single environment. In each new excursion, a new environment is introduced hence signifying a new perceptual cycle. However, the actor-critic module changes during a *developmental time scale* which is much slower than the perceptual one. One may even consider episodic learning, wherein each excursion is a single episode that changes the actor-critic module. Hence, the policies change little every excursion, but change drastically along development in which many environments are encountered. For this reason (as well as their local nature) the policies in our framework *converge* with time to fixed motor primitives.

In contrast, in most other architectures (Oudeyer et al. 2007; Tishby and Polani 2011; Der and Martius 2012; Pape et al. 2012; Little and Sommer 2013) ((Der and Martius 2012) representing the other extreme), the "perceptual" and "developmental" time-scales are identical, such that during a single encounter with a novel environment, policies continuously change and never converge. They are always seeking novel areas in state-action spaces. To accommodate this ever-changing policy, in our framework we introduce a pool of policies, namely, the hierarchical curiosity loops. Thus, during exploration of a single environment, where other models have a single ever-changing policy, we have several converged exploration primitives.

6.3 Maximizing information gain versus novelty management

There are several models of intrinsic motivation, or artificial curiosity, that attempt to maximize or minimize some form of information gain or surprise (Schmidhuber 1990; Polani 2009; Schmidhuber 2010; Friston 2010; Tishby and Polani 2011; Der and Martius 2012; Little and Sommer 2013). However, there was no attempt to explain the rich repertoire of observed and quantitatively analyzed behaviors in multiple scenarios. We believe that animals do not in fact try to optimize information gain alone. As our framework suggests, animals display a delicate interplay between maximizing information gain and reducing novelty. The former behavior can probably be explained by alternative models, such as those discussed in Refs. (Polani 2009; Schmidhuber 2010; Friston 2010; Tishby and Polani 2011). Furthermore, the retreat behavior is an adaptive yet pre-programmed behavior, very similar to moving down a novelty gradient to a specific location. One novel aspect of this contribution is the hypothesis that animals switch between both via a novelty-management principle and that some observed behaviors that were previously thought to be of one source, can actually be explained by a combination of novelty seeking and avoidance.

The rationale behind the novelty-aversive principle is the observed tendency of animals to become agitated whenever presented with too much novelty (Barnett 1958; Misslin and Cigrang 1986; File 2001; Hughes 2007; Fonio et al. 2009) and return to a known location. This tendency represents a delicate trade-off between the curiosity drive, which seeks novelty, and anxiety that stems from too much novelty.

Furthermore, the retreat primitive is closely related to the concept of safety and ergodicity (Moldovan and Abbeel 2012), which states that a safe policy is one that guarantees that all visited states can be reached. In contrast, a non-ergodic and unsafe policy can venture into states that do not allow a probable return to other previously visited states, e.g. going down an unclimbable pit. In (Moldovan and Abbeel 2012) a similar retreat policy is found and guarantees the safety of another sought-out policy, by constantly checking that the developed policy can always return back. In our context, the retreat motor primitive guarantees that the agent will always be able to return to its base-state. One

can thus suggest that animals try to maximize *useful information* (Tishby and Polani 2011), where usefulness in this context is their safety.

Moreover, in contrast to the approach-avoidance theory (Elliot 2006), which can view the curiosity-driven primitives as approach and the retreat primitive as avoidance, our framework has a single objective, i.e. novelty control. In other words, there are no two competitive goals, whose balance brings about this description of the behavior, but rather a single balancing adaptive goal with a single variable, namely, novelty. We view this difference as an advantage, since it can explain the same behavior with less variables.

6.4 Single complex behavior versus hierarchical architecture

There is a growing body of research that focuses on learning of motor primitives and acquisition of skills (Barto et al. 2004; Pape et al. 2012). Furthermore, in the field of reinforcement learning (RL) the appearance of hierarchical reinforcement learning (HRL) (Barto and Mahadevan 2003; Sutton et al. 2011) has instigated a new paradigm wherein large and complex tasks can be automatically decomposed to smaller tasks, each solved via an RL algorithm. In our context, this would have surmounted to decomposing the exploration task into smaller subtasks, each resulting in an emergent behavior. A similar approach was taken by Oudeyer, wherein the entire large state-action space is slowly segmented to smaller regions, each representing a different context, wherein a different actor is learned (Oudeyer et al. 2007).

We have chosen a different approach by deliberately reducing the actor's state space to those that can be easily transferable to other yet similar environments, similar to the concept of agent-space options (Konidaris and Barto 2007). The transference can be made only within the same modality and same "type" of environment, in contrast to higher level skill transference. By restricting the actors to be local and invariant, we created a situation in which the exploration task cannot be solved by a single curiosity loop. We have introduced the hierarchical construction of multiple curiosity loops, each starting where the previous finished, thus autonomously discovering features with decreasing novelty content. Constant novelty is not maintained via a global optimizer that attempts to keep it constant (Saig et al. 2012), but rather by a delicate balance between the curiosity loops and the novelty-aversive retreat policy. In contrast to Oudeyer and similar approaches, the separation to "contexts" is not done in the state-action space, but rather in "novelty-space", wherein each feature with different novelty has its own optimized exploration behavior.

Furthermore, the emergent behaviors can be view as acquired skills of exploration (Barto et al. 2004; Pape et al. 2012). A single policy that attempts to explore the entire space by adapting according to the novelty input will be continuously changing (Der and Martius 2012; Schmidhuber 2010; Ngo et al. 2012) and will never converge. Our model, on the other hand, attempts to discover the motor primitives of exploration, i.e. given that a specific type of environment is encountered many times, but with different "parameters", what is the optimal policy to explore it? There should be a single (or hierarchical construction of multiple) *stationary policies*. In order to discover them, the curiosity loop with information gain intrinsic reward is employed.

Furthermore, in contrast to (Gordon and Ahissar 2012) in which each curiosity loop had its own perceiver, here there is a single perceiver for all the curiosity loops. This difference is due to different motivations: in (Gordon and Ahissar 2012) we were interested in comparing actors that converged due to different perceivers, e.g. forward and inverse models, and the state-space itself changed with each hierarchical level. In this contribution, we are interested in a single perceiver that contains the entire exploration space; however, the critics and actors are local and cannot be used to learn the entire exploration space, thus requiring the hierarchical construction, even with a single perceiver. The two architectures are not mutually exclusive, namely, one can have several hierarchical curiosity loop architectures, each for a different perceiver, wherein the critics and actors are local and invariant.

6.5 External versus internal goals

The formal framework presented here deals only with intrinsic motivation and reward. Hence, the behavioral scenarios considered are limited to exploration of novel environments. However, animals (and robots) perform goal-directed actions as well and the relation between the two must also be considered (Tishby and Polani 2011). In our opinion, during development, convergence of exploration motor primitives serves as a major process by which motor primitives emerge at large. Hence, when attempting to construct a complex goal-directed motor strategy, the agent has as its disposal a pool of motor primitives, from the exploration scenarios. Thus, for example, one can hypothesize that known motor primitives (Flash and Hochner 2005) are learned during exploratory-guided behavior and not a goal-directed one.

In a more general sense, during ontogenetic development, one can consider many seemingly goal-directed behaviors, usually thought of as those having an external reward, as exploratory behaviors, generating intrinsic reward. For example, reaching for a toy can be thought of as exploration of hand-object interaction, hence exploratory, and not

necessarily a goal-directed behavior, as designed by many experimenters. Thus, a motor primitive converged during such behaviors comply with our definition of exploratory motor primitives. During adult behavior, truly goal-directed complex motions can be composed of these motor primitives.

7 Discussion

We presented a single ubiquitous developmental formal framework which learns exploratory behavior by repeated interactions with the environment which results in the convergence of motor primitives of exploration, each efficient in learning a specific feature of the environment. The framework is augmented with a mechanism that preserves a delicate balance between exploration and novelty-avoidance, thus maintaining a constant level of novelty. The emergent behavior exhibits exploration of increased complexity and dimensionality, interspersed with frequent returns to a place of low-novelty.

7.1 Neurophysiological and neuroanatomical considerations

The framework architecture suggests a neuroanatomical and neurophysiological mapping of the components onto specific brain structures: The perceiver learns to predict future sensory information based on the current motor-sensory state. Two structures form the primary candidates for this function: The cerebellum and the thalamocortical system. In the whisker system, the preferred candidate for implementing the perceiver is the thalamocortical system, which contains the updated information about the on-going motor-sensory state (Guillery and Sherman 2012; Yu et al. 2013) and is capable of predicting future states (Ahissar 1998; Ahissar and Oram 2013). Candidate structures for the actors are motor nuclei, primarily the facial nucleus, pre-motor brainstem nuclei and mid-brain sensorimotor nuclei (Ahissar and Kleinfeld 2003; Kleinfeld et al. 2006; Diamond et al. 2008; Ahissar and Knutsen 2008).

The novelty management unit suggests a centralized region that receives input of novelty and switches between the novelty seeking and novelty aversive behaviors. However, an action selection mechanism can serve just as well, as long as its reward input is intrinsic reward. Hence, the basal ganglia is a good candidate for the location for novelty management processing (Redgrave 2007). Furthermore, the recent discovery of two complementary paths, one for reward and one for punishment (Sesack and Grace 2009) is a strong support for the framework, which distinguished categorically between the two behaviors. This is in contrast to a single behavior that tries to maintain a constant novelty, which would not necessitate two separate mechanisms for raising and lowering novelty.

Novelty aversive behavior is suggested to be tightly connected to fear-related regions due to the hypothesis that it relates to anxiety (Misslin and Cigrang 1986; Fonio et al. 2009). However, according to our formal framework it should be tightly related to perceptual learning regions, e.g. place cells in the hippocampus in an arena exploration scenario, since they must be accessible when determining the next action. In other words, novelty aversive behavior is similar to goal-directed policy, whose goal is to return to a known safe state. This requires already-learned information about the environment and is to be contrasted to other fear-related behavior such as freezing, which does not.

Considering the basic curiosity loop, the framework predicts novel neural circuitry during development. In order to facilitate rewarding information gain, there should be a strong *input* connectivity to the rewarding system from internal model areas, e.g. cerebellum (Shadmehr and Krakauer 2008; Lalazar and Vaadia 2008) and sensory perception areas, e.g. primary sensory cortices (Matyas et al. 2010; Feldmeyer et al. 2012; Bastos et al. 2012). The framework predicts that this connectivity should be stronger during development to allow convergence of the stereotypical exploratory behaviors apparent in adult rats. Furthermore, the conveyed information in these connections should code prediction error signals (Shadmehr and Krakauer 2008; Lalazar and Vaadia 2008).

Within each curiosity loop there are several internal variables that play critical roles in the framework. The first is the average reward, which determines the novelty threshold of that loop, i.e. novelty greater than it instigates retreat. It was suggested that average reward is related to opportunity costs and latency between action switching (Niv et al. 2007; Cools et al. 2011). Furthermore, it was suggested to be related to tonic dopamine in nucleus accumbens. Similarly, in our framework, the average reward serves as the reference point which determines switching between different motor primitives.

7.2 Development versus evolution

The formal framework we propose is based on ontogenetic development via experience of the individual's interaction with the world and the resulting intrinsic reward it receives. The end result is a behavioral policy composed of the interplay between the learned exploration primitives and the retreat primitive. However, similar considerations could be made regarding the phylogenetic development, where fitness and reproduction rate substitute intrinsic rewards. In other words, neural circuits of exploration primitives could be developed by evolution and natural selection, if efficient learning of the environment increases survival rate.

One can further speculate that simpler yet efficient exploration primitives, e.g. periodic whisking, would be "found" by evolutionary mechanisms and coded in the experience-independent neural circuitry development. This may explain the existence of central-pattern-generators as part of the exploratory behaviors; an evolutionary recruitment of a "hard-wired" circuitry for efficient exploration (Deschenes et al. 2012).

This phylogenetic argument does not contradict the proposed model, but rather suggests a complementary, longer time-scale adaptation mechanism to common environmental features. Thus, for example, experiments which alter environments experienced by pups can help separate which exploration primitives emerge via intrinsic reward and which are innate. More specifically, it promotes the testable prediction that whisking patterns predicted by the framework can be changed via introduction of different rearing environments.

7.3 Novelty-controlled robots

While robots do not experience fear (yet), their designers and human companions do. Hence, integrating neophobia into a robot may serve the purpose of keeping the robot intact and its surrounding safe (Moldovan and Abbeel 2012). In this form, keeping novelty at a constant rate restricts the robot and its handler of "taking chances" and exploring the entire state-action space immediately. The robot still acts according to the curiosity drive, but is balanced by not trying too many new things at once. In an environment with delicate features, such as human companions, this may be an important prerequisite. Novelty control and exploration primitives can thus be another example of the benefits of incorporating animal behavior into robots.

To conclude, we have presented a general framework which puts forth the notion that novelty is managed by exploring animals, rather than maximized or minimized. The framework presents the emergence of exploration motor primitives, which optimize exploration of specific features of the perceived environment. The features also emerge in a hierarchical manner due to the interaction with the environment. Exploration behavior is shown to be the result of an intricate interplay between exploration motor primitives and retreat, thus balancing novelty input. We suggest plausible neural substrates for each of framework components, thus allowing several novel behavioral, neuroanatomical and neurophysiological predictions.

Acknowledgments We thank Karl Friston, Daniel Polani and Fritz Sommer for helpful comments and discussions. This work was supported by the Israel Science Foundation grant #749/10 and the United States-Israel Bi-national Science Foundation grant #2011432. E.A. holds the Helen Diller Family Chair in Neurobiology.

Conflict of interests The authors declare that they have no conflict of interest.

References

Ahissar, E. (1998). Temporal-code to rate-code conversion by neuronal phase-locked loops. *Neural Computer*, *10*(3), 597–650.

Ahissar, E., & Kleinfeld, D. (2003). Closed-loop neuronal computations: focus on vibrissa somatosensation in rat. *Cereb Cortex*, *13*(1), 53–62.

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

Ahissar, E., & Oram, T. (2013). Thalamic relay or cortico-thalamic processing? Old question, New Answers. Cerebral Cortex: bht296.

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.

Baldassarre, G. (2011). What are intrinsic motivations? a biological perspective. *IEEE International conference developmental learning (ICDL)*, (Vol. 2, pp. 1–8).

Barnett, S.A. (1958). Exploratory behaviour. *Br J Psychol*, *49*(4), 289–310.

Barto, A.G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamical System*, *13*(1–2), 41–77.

Barto, A.G., Singh, S., Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *International conference on developmental learning (ICDL)*.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Behera, L., Gopal, M., Chaudhury, S. (1995). Self-organizing neural networks for learning inverse dynamics of robot manipulator. *In IEEE/IAS International conference on industrial automation and control (I A & C'95)* (pp. 457–460).

Berg, R.W., & Kleinfeld, D. (2003). Rhythmic whisking by rat: Retraction as well as protraction of the vibrissae is under active muscular control. *Journal of Neurophysiol*, *89*(1), 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).

Cools, R., Nakamura, K., Daw, N.D. (2011). Serotonin and dopamine: Unifying affective, activational, and decision functions. *Neuropsychopharmacology*, *36*(1), 98–113.

Der, R., & Martius, G. (2012). *The playful machine. Cognitive System Monographia*. Springer.

Deschenes, M., Moore, J.W., Kleinfeld, D. (2012). Sniffing and whisking in rodents. *Current Opinion in Neurobiology*, *22*(2), 243–250.

Deutsch, D., Pietr, M., Knutsen, P.M., Ahissar, E., Schneidman, E. (2012). Fast feedback in active sensing: touch-induced changes to whisker-object interaction. *PLoS One*, *7*(9), e44, 272.

Diamond, M.E., von Heimendahl, M., Knutsen, P.M., Kleinfeld, D., Ahissar, E. (2008). Where and what in the whisker sensorimotor system. *Natural Reviews Neuroscience*, *9*(8), 601–612.

Elliot, A.J. (2006). The hierarchical model of approach-avoidance motivation. *Motivaton and Emotion*, *30*, 111–116.

Fanselow, E.E., Sameshima, K., Baccala, L.A., Nicolelis, M.A. (2001). Thalamic bursting in rats during different awake behavioral states. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(26), 15330–5.

Feldmeyer, D., Brecht, M., Helmchen, F., Petersen, C.CH., Poulet, J.FA., Staiger, J.F., Luhmann, H.J., Schwarz, C. (2012). Barrel cortex function. *Progress in Neurobiology*, *103*(0), 3–27.

File, S.E. (2001). Factors controlling measures of anxiety and responses to novelty in the mouse. *Behavioural Brain Research*, *125*(1–2), 151–7.

Flash, T., & Handzel, A.A. (2007). Affine differential geometry analysis of human arm movements. *Biological Cybernetics*, *96*(6), 577–601.

Flash, T., & Hochner, B. (2005). Motor primitives in vertebrates and invertebrates. *Current Opinion in Neurobiology*, *15*(6), 660–6.

Fonio, E., Benjamini, Y., Golani, I. (2009). Freedom of movement and the stability of its unfolding in free exploration of mice. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21, 335–40.

Fox, C.J., Girdhar, N., Gurney, K.N. (2008). A causal bayesian network view of reinforcement learning. In Twenty-first international florida artificial intelligence research society conference (pp. 109–110). AAAI Press.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature Reviews Neuroscience*, *11*(2), 127–38.

Frommberger, L., & Wolter, D. (2010). Structural knowledge transfer by spatial abstraction for reinforcement learning agents. *Adaptive Behavior - Animals, Animats, Software Agents, Robot, Adaptive System*, *18*(6), 507–525.

Gao, P., Bermejo, R., Zeigler, H.P. (2001). Whisker deafferentation and rodent whisking patterns: Behavioral evidence for a central pattern generator. *Journal of Neuroscience*, *21*(14), 5374–5380.

Gordon, G., & Ahissar, E. (2012). Hierarchical curiosity loops and active sensing. *Neural Network*, *32*, 119–29.

Gordon, G., Kaplan, D.M., Lankow, B., Little, D.Y., Sherwin, J., Suter, B.A., Thaler, L. (2011). Toward an integrated approach to perception and action: conference report and future directions. *Frontiers System Neuroscience*, *5*, 20.

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. *Journal of Neurophysiology*, *101*(2), 862–74.

Grant, R.A., Mitchinson, B., Prescott, T.J. (2012). The development of whisker control in rats in relation to locomotion. *Developmental Psychobiology*, *54*(2), 151–168.

Guillery, R.W., & Sherman, S.M. (2012). The thalamus as a monitor of motor outputs. *Philosophical Transactions R Society London B Biological Sciences*, *357*(1428), 1809–1821.

Harish, O., & Golomb, D. (2010). Control of the firing patterns of vibrissa motoneurons by modulatory and phasic synaptic inputs: a modeling study. *Journal of Neurophysiology*, *103*(5), 2684–99.

Harlow, H.F. (1950). Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *Journal of Comparative & Physiological Psychology*, *43*(4), 289–94.

Hill, D.N., Bermejo, R., Zeigler, H.P., Kleinfeld, D. (2008). Biomechanics of the vibrissa motor plant in rat: Rhythmic whisking consists of triphasic neuromuscular activity. *Journal of Neurophysiology*, *28*(13), 3438–3455.

Hughes, R.N. (2007). Neotic preferences in laboratory rodents: issues, assessment and substrates. *Neuroscience and Biobehavioral*, *31*(3), 441–64.

Kaelbling, L.P., Littman, M.L., Moore, A.W. (1996). Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kawato, M.M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, *9*, 718–727.

Kleinfeld, D., Ahissar, E., Diamond, M.E. (2006). Active sensation: insights from the rodent vibrissa sensorimotor system. *Current Opinion in Neurobiology*, *16*(4), 435–44.

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*(1), 35–42.

Konidaris, G., & Barto, A. (2007). Building portable options: skill transfer in reinforcement learning. In *Proceedings of the 20th international joint conference on artifical intelligence* (pp. 895–900). Hyderabad: Morgan Kaufmann.

Lalazar, H., & Vaadia, E. (2008). Neural basis of sensorimotor learning: modifying internal models. *Current Opinion in Neurobiology*, *18*((6)), 573–581.

Leiser, S.C., & Moxon, K.A. (2007). Responses of trigeminal ganglion neurons during natural whisking behaviors in the awake rat. *Neuron*, *53*(1), 117–33.

Little, D.Y., & Sommer, F.T. (2013). Learning and exploration in action-perception loops. Frontiers in Neural Circuits (in press).

Matyas, F., Sreenivasan, V., Marbach, F., Wacongne, C., Barsy, B., Mateo, C., Aronoff, R., Petersen, C.C. (2010). Motor control by sensory cortex. *Science*, *330*(6008), 1240–3.

Misslin, R., & Cigrang, M. (1986). Does neophobia necessarily imply fear or anxiety?. *Behavior Processes*, *12*(1), 45–50.

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. *Proceedings of the Biological Sciences*, *274*(1613), 1035–41.

Miyazaki, M., Yamamoto, S., Uchida, S., Kitazawa, S. (2006). Bayesian calibration of simultaneity in tactile temporal order judgment. *Nature Neuroscience*, *9*(7), 875–7.

Moldovan, T.M., & Abbeel, P. (2012). Safe exploration in markov decision processes. In *ICML 2012*.

Ngo, H., Luciw, M., Foerster, A., Schmidhuber, J. (2012). Learning skills from play: Artificial curiosity on a katana robot arm. In *IJCNN 2012*.

Nicolelis, M.A., Baccala, L.A., Lin, R.C., Chapin, J.K. (1995). Sensorimotor encoding by synchronous neural ensemble activity at multiple levels of the somatosensory system. *Science*, *268*(5215), 1353–8.

Niv, Y., Daw, N.D., Joel, D., Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacol (Berl)*, *191*(3), 507–520.

Oudeyer, P.Y., Kaplan, F., Hafner, V.V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computations*, *11*(2), 265–286.

Ouyang, P.R., Zhang, W.J., Gupta, M.M. (2006). An adaptive switching learning control method for trajectory tracking of robot manipulators. *Metchatronics*, *16*, 51–61.

Pape, L., Oddo, C.M., Controzzi, M., Cipriani, C., Frster, A., Carrozza, M.C., Schmidhuber, J. (2012). Learning tactile skills through curious exploration. *Frontiers in Neurorobotics*, 6.

Polani, D. (2009). Information: currency of life. *HFSP Journal*, *3*(5), 307–16.

Precup, D., Sutton, R.A., Dasgupta, S. (2001). Off-policy temporal difference learning with function approximation. In *Proceedings of the eighteenth international conference on machine learning* (pp. 417–424).

Redgrave, P. (2007). Basal ganglia. *Scholarpedia*, *2*(6), 1825.

Richardson, M.J., & Flash, T. (2002). Comparing smooth arm movements with the two-thirds power law and the related segmented-control hypothesis. *Journal of Neuroscience*, *22*(18), 8201–11.

Saig, A., Gordon, G., Assa, E., Arieli, A., Ahissar, E. (2012). Motor-sensory confluence in tactile perception. *Journal of Neuroscience*, *32*(40), 14,022–32.

Schmidhuber, J. (1990). A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the first international conference on simulation of adaptive behavior on from animals to animats* (Vol. 116542, pp. 222? 227). MIT Press.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230–247.

Schultz, W., Dayan, P., Montague, P.R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–9.

Semba, K., Szechtman, H., Komisaruk, B.R. (1980). Synchrony among rhythmical facial tremor, neocortical 'alpha' waves, and thalamic non-sensory neuronal bursts in intact awake rats. *Brain Research*, *195*(2), 281–98.

Sesack, S.R., & Grace, A.A. (2009). Cortico-basal ganglia reward network: Microcircuitry. *Neuropsychopharmacology*, *35*(1), 27–47.

Shadmehr, R., & Krakauer, J.W. (2008). A computational neuroanatomy for motor control. *Experimentalis Brain Research*, *185*(3), 359–81.

Simony, E., Bagdasarian, K., Herfst, L., Brecht, M., Ahissar, E., Golomb, D. (2010). Temporal and spatial characteristics of vibrissa responses to motor commands. *Journal of Neuroscience*, *30*(26), 8935–8952.

Singh, S., Lewis, R.L., Barto, A.G., Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions Autonomous Mental Development*, *2*(2), 70–82.

Stolle, M., & Precup, D. (2002). In *Learning options in reinforcement learning, lecture notes in computer science* (Vol. 2371, pp. 212–223). Berlin Heidelberg: Springer.

Sutton, R., Precup, D., Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence in Engineering*, *112*, 181–211.

Sutton, R.A., Modayil, J., Delp, M., Degris, T., Pilarski, P.M., White, A., Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th international conference on autonomous agents and multiagent systems - volume 2, international foundation for autonomous agents and multiagent systems*, 2031726 (pp. 761–768).

Sutton, R.S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44.

Szwed, M., Bagdasarian, K., Ahissar, E. (2003). Encoding of vibrissal active touch. *Neuron*, *40*(3), 621–30.

Szwed, M., Bagdasarian, K., Blumenfeld, B., Barak, O., Derdikman, D., Ahissar, E. (2006). Responses of trigeminal ganglion neurons to the radial distance of contact during active vibrissal touch. *Journal of Neurophysiology*, *95*(2), 791–802.

Tchernichovski, O., & Benjamini, Y. (1998). The dynamics of long-term exploration in the rat. part ii. an analytical model of the kinematic structure of rat exploratory behavior. *Biological Cybernetics*, *78*(6), 433–40.

Tchernichovski, O., Benjamini, Y., Golani, I. (1998). The dynamics of long-term exploration in the rat. part i. a phase-plane analysis of the relationship between location and velocity. *Biological Cybernetics*, *78*(6), 423–32.

Tinbergen, N. (1951). *The study of instinct*. New York: Oxford University Press.

Tishby, N., & Polani, D. (2011). *Information theory of decisions and actions. Springer series in cognitive and neural systems* (chap. 19 pp. 601–636). New York: Springer.

Towal, R.B., & Hartmann, M.J. (2006). Right-left asymmetries in the whisking behavior of rats anticipate head movements. *Journal of Neuroscience*, *26*(34), 8838–46.

Towal, R.B., & Hartmann, M.J. (2008). Variability in velocity profiles during free-air whisking behavior of unrestrained rats. *Journal of Neurophysiology*, *100*(2), 740–52.

Vergassola, M., Villermaux, E., Shraiman, B.I. (2007). Infotaxis as a strategy for searching without gradients. *Natural*, *445*(7126), 406–9.

Wawrzynski, P., & Pacut, A. (2004). Model-free off-policy reinforcement learning in continuous environment. In *Proceedings of the 2004 IEEE international joint conference on neural networks, 2004.* (vol 2, pp. 1091–1096).

Weng, J. (2004). Developmental robotics: theory and experiments. *International Journal Humanoid Robotics*, *1*(2), 199–236.

Whishaw, I.Q., Gharbawie, O.A., Clark, B.J., Lehmann, H. (2006). The exploratory behavior of rats in an open environment optimizes security. *Behavior Brain Research*, *171*(2), 230–9.

Yu, C., Horev, G., Rubin, N., Derdikman, D., Haidarliu, S., Ahissar, E. (2013). Coding of object location in the vibrissal thalamocortical system. *Cerebral Cortex: bht241*.