About the constuctivist role of self-evaluation for interactive learnings and self-development

C. Giovannangeli
CNRS UMR8051 ETIS
Neurocybernetic Team
Cergy-Pontoise University - ENSEA
2, avenue Adolphe-Chauvin
95302 Cergy-Pontoise, France
christophe.giovannangeli@gmail.com

S. Boucenna
CNRS UMR8051 ETIS
Neurocybernetic Team
Cergy-Pontoise University - ENSEA
2, avenue Adolphe-Chauvin
95302 Cergy-Pontoise, France

P. Gaussier
CNRS UMR8051 ETIS
Neurocybernetic Team
Cergy-Pontoise University - ENSEA
2, avenue Adolphe-Chauvin
95302 Cergy-Pontoise, France
Member of the Institut Universitaire de France

I. Introduction

A. About the role of self-evaluation in robotics

The proposed work raises the problem of autonomous learnings in robotics. Whatever the morphology of the robots and the various skills they could acquire according to their morphology, robots need to be able to self-evaluate, not only in order to guide their autonomous development trough the vast sensory-motor space, but also in order to verify that previous learning are still pertinent and to readapt their knowledge when previous learnings becomes erroneous. Such capabilities appear crucial in developmental robotics and should largely enrich the possible manifolds of social and physical interactions in which robots could be involved. As a support to this assumption, the presentation will focus on a mature bio-inspired neural network architecture, which uses an autonomous, online, and interactive learning, allowing a robot to achieve sensory-motor tasks and planning, in the frame of navigation. The role of self-evaluation for interactive learnings and self-development is highlighted.

B. Interactive learning of navigation tasks: a mature bio-inspired architecture

For the last decade, we have developed a neural network architecture, inspired from the brain functioning, efficient for visual navigation and imitation tasks. This architecture involves a model of the visual system, the temporal and parietal cortices also called the *what and where* pathways, the hippocampus, the prefrontal cortex, the basal ganglia, the motor cortex and the cerebellum [1]. The use of a robot standing for a simulation of an animal or a human provides an efficient mean to validate our models and verify if the global dynamics of the robot/environment interactions corresponds to those observed by the neurobiologists and the psychologists. As visual navigation is concerned, our model enables a robot to learn visual landmarks, to associate them to their spatial properties (azimuth and elevation) in order to build a constellation of landmarks (a set of triplet *landmark-azimuth-elevation*). The activity of the neurons recognizing such a constellation is similar to the activity of the large place cells recorded in the entorhinal cortex of the rat [19]. A set of *place-action* associations, learned in one-shot and adapted online, creates an attraction basin enabling our robot to go back to a learned location or to follow an arbitrary visual path in indoor environments [4]. The robustness of our visual place cells has recently been optimized for outdoor environments [9], enabling the robot to achieve sensory-motor tasks in indoor as well as in large outdoor environments with a low computation load [7] (see fig. 1). The behavior is also robust to kidnapping, to object and landmark addition or removal, to the presence of mobile obstacles and to severe visual field occlusions[1].

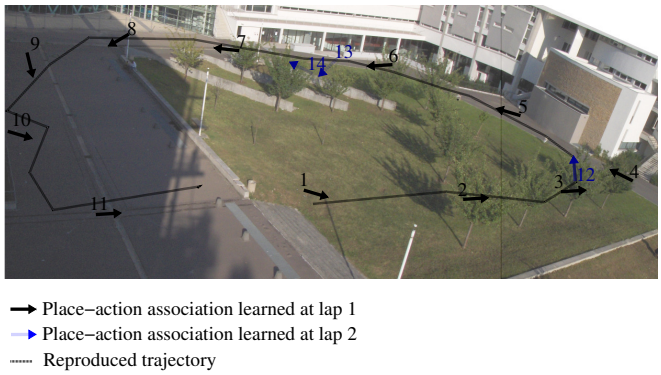

Fig. 1. Interactive learning (proscriptive teaching) of a 200 meters path. After two laps of training, the robot Pioneer AT III ActivMedia closes the loop in 9 mn. The architecture is paralleled on four processors. The speed of the robot is limited due to low level drivers of the Pioneer.

More recently, we investigate how such sensory-motor behaviors (place-action strategy) could be learned by interacting with a naive human teacher (see fig. 1) [8]. We pointed out that the richness of a real HRI (human-robot interaction), as opposed to a pre-determined guidance strategy such as a prescriptive guidance commonly referred as programming by demonstration (which is in fact far from being an interactive learning process since the behavior of the robot does not

This work was supported by the Délégation Générale pour l'Armement (DGA), contract n^o 04 51 022 00 470 27 75, the Feelix Growning project and the Institut Universitaire de France.

J.P. Banquet works in the neurocybernetic team on the neurobiological aspects of the models.

G. Désilles is our scientific correspondant from the DGA.

[1]demonstrative movies are available on the web pages of the authors.

modify the teacher behavior) or a proscriptive strategy consisting of correcting the robot when it commits errors (which is the first step toward an interactive learning, since the behavior of the robot influences the teacher behavior, modifying the robot behavior, and so on and so on...), acts as a cognitive catalyst, enhancing the precision of the reproduced behavior as well as its functioning domain (size of the attraction basin) [8]. During such a HRI, a real communication based on actions emerges [14]: the teacher communicates by acting on the joystick, and the robot communicates by behaving according to its learned sensory-motor dynamics. The interaction is composed of three different phases which alternate: prescriptive teaching phases, proscriptive teaching phases and observation/demonstration phases. The three phases alternate in time and space according to the evolution of the teaching, defining an interaction rhythm.

II. Learning progress and self-evaluation capabilities

A. Related work and motivations

However, although the learning is autonomous, the problem of evaluating the robot learning early arises in the context of a robot evolving in a human world. Reasonably, it is impossible to test all the situations the robot will be confronted to. Based on this remark, the idea that the robot should self-evaluate becomes central. An intuitive starting point for self-evaluation is to compute the learning progress. In a developmental perspective, some authors proposed to use the learning progress as a reward for a predictor, in order to guide its exploration of the sensory-motor contingencies [13], [17], [18]. The learning progress was classically computed as the derivative of the mean prediction error. More precisely, the authors proposed an "mixture of expert" architecture [22], which predicts the future sensory-motor state $(s(t+1), m(t+1), r(t+1))$, knowing $smr(t)$ (s: sensation, m: motor, r: reward). The learning progress of the sensory prediction defined as $[e_m(t-\tau^e) - e_m(t)]^+$ is then used as a reward for the SMR predictor (τ^e being the delay between the two mean errors (past error and current error), used to compute the progress). The authors observed that performing the action that maximizes the learning progress leads the robot to focus on state in which progress is possible, and to avoid unpredictable state as well as easily predictable state. The robot exhibits the capability to self-develop its skills, by analysing among all the sensory-motor associations, which are known, which are unpredictable, and especially which are sources of progress. The robot also exhibits a kind of curiosity, speeding-up its development [20], since the robot oscillate around *the frontier that separates mastered know-how from unmastered know-how* [13]. The counterpart is that the robot never try to become better in a given task but always try to reach a unmastered state.

Yet, the learning progress appears as a rich signal that could help to reliably trigger or suspend learning phases. We defined the stagnation as the fact that the predictor no longer progresses, thought it is learning. In such a case, learning can be suspended and an evaluation measure of the predictor can be estimated. As long as the accuracy remains constant, the system does not need to learn. Changes in the accuracy of the prediction are likely to be induced by a change in the physical world, corresponding to novelty, and suggesting that learning can be necessary.

Let us focus on the computation of the learning progress. The following concepts can be applied to all prediction machines, able to generate an error $e(t)$ between the prediction and the reality (the mean error is noted $\epsilon(t) = \bar{e}(t) = \frac{1}{\tau}\sum_{i=0}^{\tau-1} e(t-i)$). The learning progress is generally defined as:

$$Prog(t) = \Big[-\frac{\Delta\epsilon(t)}{\Delta t}\Big]^+$$

corresponding to the decrease of the mean static error. The progress is positive when the current mean error is lower than the past mean error. Otherwise, progress is null. Even if this formulation of the progress is correct, it is not complete as it will be explained.

A first observation is that error can reduce even in absence of adaptation. This is not frequent, but this can happen. This corresponds to the fact the real phenomenon converges toward the learned phenomenon: occurrence of a progress which does not result from the learning convergence, for example when a teacher adapts its behavior to the difficulties of its student (the error is decreasing because the difficulty is decreasing). Except this particular case, since each predictor has a built-in accuracy for the phenomenon it predicts, the predictor should stabilize to provide a given mean error with a given standard deviation. In this case, whatever the phenomenon is, the progress and its mean value should become null and oscillate around 0 after a given time. Otherwise, it would correspond to a continuous progress which is only possible in theoretical cases in which the signal to predict is constant. Stabilization of the error is called "stagnation". Even if the predictor is not able to correctly predict the phenomenon, we consider the predictor stagnates when it is habituated[2] to provide a particular error (ie: when its mean progress has fallen under 0).

For a stagnating predictor, changes in the nature of the phenomenon (which is called here "novelty") make the predictor produce a higher error than the current error (changes that decrease the error has already been discussed). The former definition of the progress as the positive decrease of the mean error does not take into account that predictions can become erroneous. Indeed, if the error increases due to a novelty, the progress will become negative, whereas it should become positive: reasonably, a novelty is likely to imply something to learn, hence a positive progress. After an erroneous prediction, even if the system is learning, the progress as defined earlier will become negative during a period that depends on τ and τ^e. Yet, it seems legitimate to consider that the novelty detection should have a direct impact of the estimated progress value.

[2]We use the term "stagnation rather than "habituation" to not create a confusion with the classical synaptic habituation.

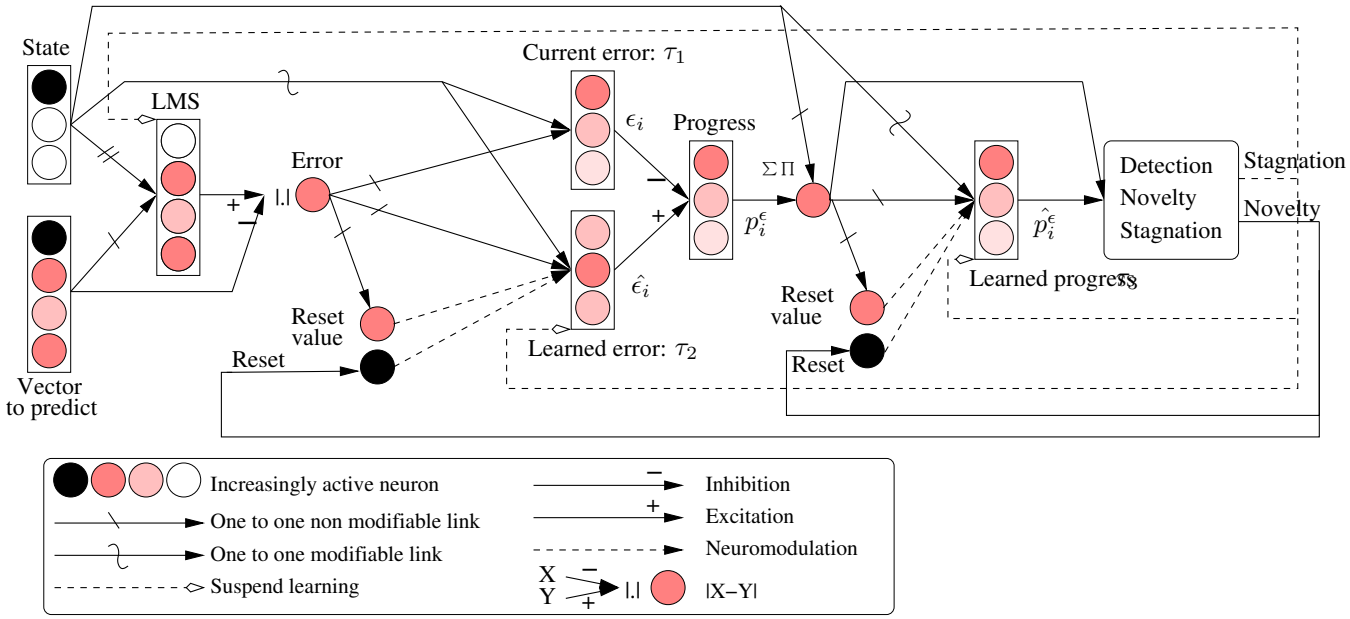

Fig. 2. Architecture for self-evaluation. The progress is computed as the difference between the learned error and the current error. The stagnation is detected when the learned progress becomes negative which suspend the learning. If an abnormal error occurs, the progress changes and novelty is detected. When novelty is detected, a signal allows to reset the learned error and the learned progress to the current error and progress value. The computed progress becomes positive and the stagnation detector is inhibited. A new learning phase is triggered

B. *Interaction progress/stagnation/novelty*

Aware of these limitations, we propose a more precise computational definition of the progress, illustrated by the architecture of fig. 2: the progress is the difference between the learned error $\hat{\epsilon}(t)$ and the current error $\epsilon(t)$. A positive progress p_k^ϵ corresponds to the fact that the mean error $\epsilon_k(t)$ of the predictor for the state k is lower than the learned value $\hat{\epsilon_k}(t)$:

$$p_k^\epsilon(t) = \hat{\epsilon_k}(t) - \epsilon_k(t)$$

We assume here that when the learning is stabilized (corresponding to stagnation phases), the mean progress will fall under 0. Practically, when the learning has converged, $\epsilon_k(t)$ stagnates, $\hat{\epsilon_k}(t)$ converges toward $E[\epsilon_k(t)]$, and the progress becomes negative before oscillating around 0 ($E[X]$ and V[X] are the expected value and the variance of X). The progress appears as the centered version $\gamma(0, \sigma^2_{\epsilon_k})$ of the mean error $\gamma(\nu_{\epsilon_k}, \sigma^2_{\epsilon_k})$. Indeed, p_k^ϵ oscillate around 0 with a given standard deviation:

$$\begin{aligned} V[p_k^\epsilon] &= E\Big[(E[p_k^\epsilon] - p_k^\epsilon)^2\Big] && (1)\\ &= E\Big[(p_k^\epsilon)^2\Big] && (2)\\ &= E\Big[(E[\epsilon_k] - \epsilon_k)^2\Big] = V[\epsilon_k] && (3) \end{aligned}$$

We deduce as foreseen that $\sigma^2_{p_k^\epsilon} = \sigma^2_{\epsilon_k}$. Between the instant when $p_k^\epsilon(t)$ becomes negative and the instant when $\hat{p_k^\epsilon}$ becomes negative ($\hat{p_k^\epsilon}$ is the learned version of p_k^ϵ and converges toward $E[p_k^\epsilon] = 0$), the duration is long enough for the estimator of the progress deviation (equivalent to the estimator of the error deviation) to converge. The predictor is said to stagnate, inducing the suspending of the learning (of the predictor, of the mean error and of the progress), in order to estimate in an invariant manner the predictor accuracy: during stagnation, the mean error $\hat{\epsilon_k}$ (which is fixed) and the deviation $\sigma^2_{\epsilon_k} = \sigma^2_{p_k^\epsilon}$ (which is still online computed) represents an interesting signals to measure the accuracy of the predictor. When $\hat{p_k^\epsilon}$ becomes negative, a first estimation of the progress deviation is available. Hence, novelty can be detected when a prediction generates a non conform error (the notion of conformity is the one classically used in statistic). Since the law followed by the error sample is unknown, the conformity of a sample will be accepted if this sample is close enough from its mean value. The notion of proximity to the mean value classically depends on the estimated deviation: here, we impose $|p_k^\epsilon| < 3.\sigma_{p_k^\epsilon}$:

If a novelty is detected by the non-conformity of the error sample, it is legitimate to consider that the best estimation of the learned error is the current error. The role of the novelty detection is crucial in our definition since the novelty detection triggers the reset of the learned error (which is no longer pertinent) to the current error. This reset produces a strongly positive progress, inciting the system to trigger a new learning. A novelty also suggests that the learned progress is erroneous, and that its best estimation is the current progress value. Indeed, the stagnation detector is directly inhibited by the reset of the learned progress. A learning phase can then occur.

By an active analysis of the learning progress of a predictor, it seems theoretically possible to detect stagnation and novelty, in order to meta-control this predictor, by triggering or suspending the learning phases. The whole system is able to evaluate the pertinence of its learning, and to invalidate

the learned associations when they are no longer pertinent. In the presentation, some experiments of interactive learning in a simulated environment will be presented, in order to demonstrate the expected mechanics of the proposed meta-controller. We will focus on the role of such a meta-controller in the context of the interactive learning of a sensory-motor task (ie: a navigation task, based on a place-action strategy).

C. Discussion in the field of neurobiology

Rather than addressing the problem of modeling particular structures which could be involved in the self-evaluation capabilities, we rather described a minimal set of states and transitions between these state (progress, stagnation and self-evaluation, novelty and re-adaptation) which could explain theses capabilities of animals as well as in robots. Reasonably, the learning, the maintenance, and the use of a skill implies neural loops and information transfers spread among almost all the brain structure (imagine for example how a 20 years old soccer player, in spite of its growth and changes of its muscular performance, has learned how to efficiently shoot in the ball at the end of its run). We can nevertheless introduce structures simply involved in error-prediction or reward prediction, though neurobiologists still debate on their interactions. Thus, it is commonly admitted that basal ganglia, especially the nucleus accumbens, are involved not only in reward prediction and error-reward prediction but also in the salience and valence during incentive anticipation [12]. Recently, other evidences have been given that other structures are correlated with the prediction of reward. For example, the hippocampus in able to predict the timing of a reward at a goal location [11]; reward timing is also observed in the primary visual cortex [21]; studies about the schizofrenia has shown that the orbital and dorsal prefrontal structures plays a critical role for the representation a value of outcomes and plans [10].

Emotions appear also crucial in the meta-control of the various learnings since one of its fundamental role is to express a valence associated to a prediction. The emergence of emotions within the numerous structures of the brain, and even the computational definition of what is an emotion, is far from being understood. To understand how self-evaluation takes place in the brain, the ”reward pathway” (midbrain, its projection to ventral striatum, dorsal striatum, orbito-frontal cortex and other area of the mesial prefrontal cortex) has to be better understood. We can reasonably presuppose that the computational neuroscience and bio-inspired robotics will provide, for the ongoing research in neurobiology, an efficient tool to evaluate, confirm, and especially infirm the proposed models in order to progress in our interpretation of the brain functioning.

III. Conclusion and perspectives

In the context of the interactive learning for a navigation task, if the robot is able to self-evaluate, it can inform its teacher about its mastery of the task and/or its emotional state, in order to enrich the interaction, by using facial expression, for example [2]. In unmastered situations, the robot could use repair strategies to get back the attention of the teacher, by means of a particular behavior (oscillation, stop, looking toward the teacher ...) [15], or by means of a more understandable media as an expressive robot head [2], [16]. In mastered states, the robot could become curious by choosing to not realize the learned behavior and to disobey its teacher in order to find less mastered states in which it can still progress (a communication based on the expression of emotional states could once again be very pertinent). Indeed, this could lead the robot toward states it would not have experimented if he had performed what it had learned, or if it had perform the predictable actions imposed by its teacher.

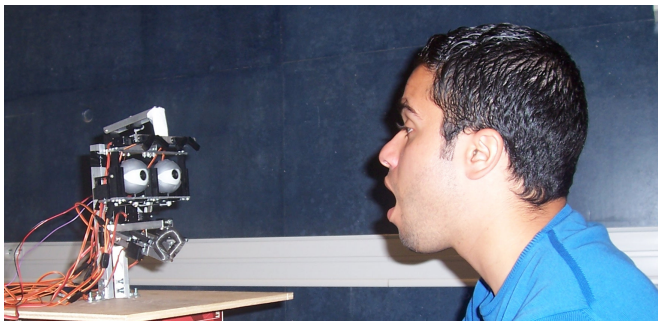

Fig. 3. S. Boucenna interacting with an emotional robotics head. The caregiver simply reproduces the facial expression of the robot, as a mother does with its baby, enabling the robot to bind its emotional internal states with expression of the caregiver. When the caregiver estimates the robot recognizes facial expressions, he manually switches the role of both agent to realize a qualitative post-analysis of the robot reaction (the robot becomes the imitator). As in navigation, if the robot is able to self-evaluate, it would autonomously switch its role and becomes the imitator. The use of self-evaluation capability seems interesting for the emergence of a natural role switching between the human and the robot.

Auto-evaluation capabilities could also help for the learning of emotional expression. The emotional binding being an extremely difficult problem, self-evaluation capabilities could be very useful to speed-up the emotional parenting between a human and a robot as illustrated by the fig. 3). In the actual state of our work, the robotics head is able to learn the emotional expressions thanks to a novelty detector, based on the rhythm of the interaction. The underlying system learns the emotional expressions if and only if this detector is constant and avoids learning if the detector fires too frequently.

Future works also will focus on the comparison of sensory-motor strategies versus planning strategies for the learning of an arbitrary path and the control of its reproduction. In our complete biological model, neurons in the hippocampus proper (CA1/CA3 regions) learn and predict transitions between successive multi-modal states [5], [3]. A cognitive map computes a latent learning of the spatial topology of the environment [23] and can be used to plan a sequence of actions to reach an arbitrary goal [6]. The influence of our progress-based meta-controller will be evaluated at all the level of this architecture. We will also study how an agent can autonomously detect it is not really doing what it aims at doing (when the robot get lost).

REFERENCES

[1] J.P. Banquet, P. Gaussier, M. Quoy, A. Revel, and Y. Burnod. A hierarchy of association in hippocampo-cortical systems: cognitive maps and navigation strategies. *Neural Computation*, 17:1339–1384, 2005.

[2] L. Canamero and P. Gaussier. *Emotion understanding: robots as tools and models*, pages 235–258. J. Nadel and D. Muir, 2005.

[3] P. Gaussier, J.-P. Banquet, F. Sargolini, C. Giovannangeli, E. Save, and B. Poucet. A model of grid cells involving extra hippocampal path integration and the hippocampal loop. *Journal of Integrative Neuroscience*, 6:447–476, 2007. in press.

[4] P. Gaussier, C. Joulain, J.P. Banquet, S. Leprêtre, and A. Revel. The visual homing problem: an example of robotics/biology cross fertilization. *Robotics and autonomous system*, 30:155–180, 2000.

[5] P. Gaussier, A. Revel, J.P. Banquet, and V. Babeau. From view cells and place cells to cognitive map learning: processing stages of the hippocampal system. *Biological Cybernetics*, 86:15–28, 2002.

[6] C. Giovannangeli and P. Gaussier. Autonomous vision-based navigation: Goal-oriented action planning by transient states prediction, cognitive map building, and sensory-motor learning. In *Proc. of the 2008 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2008)*, Nice, France, 2008.

[7] C. Giovannangeli, P. Gaussier, and G. Désilles. Robust mapless outdoor vision-based navigation. In *Proc. of the 2006 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2006)*, pages 3293–3300, Beijing, China, 2006.

[8] C. Giovannangeli and Ph. Gaussier. Human-robot interactions as a cognitive catalyst for the learning of behavioral attractors. In *16th IEEE International Symposium on Robot and Human Interactive Communication 2007*, pages 1028–1033, Jeju, South Korea, 2007.

[9] C. Giovannangeli, Ph. Gaussier, and J.-P. Banquet. Robustness of visual place cells in dynamic indoor and outdoor environment. *International Journal of Advanced Robotic Systems*, 3(2):115–124, jun 2006.

[10] J. M. Gold, J. A. Waltz, K. J. Prentice, S. E. Morris, and E. A. Heerey. Reward processing in schizophrenia: A deficit in the representation of value. *Schizophr Bu*, 2008.

[11] V. Hok, P.-P. Lenck-Santini, S. Roux, E. Save, R. U. Muller, and B. Poucet. Goal-related firing in hippocampal place cells. *Journal of Neuroscience*, 27:472–482, 2007.

[12] Cooper J.C. and Knutson B. Valence and salience contribute to nucleus accumbens activation. *Neuroimage*, 39(1):538–547, 2008.

[13] F. Kaplan and P-Y Oudeyer. Maximizing learning progress: an internal reward system for development. In *Iida, F. and Pfeifer, R. and Steels, L. and Kuniyoshi, Y., eds*. Springer-Verlag, 2004.

[14] Nathan Koenig and Maja J Mataric. Behavior-based segmentation of demonstrated task. In *IEEE International Conference on Development and Learning (ICDL)*, 2006.

[15] Claudia Muhl and Yukie Nagai. Does disturbance discourage people from communicating with a robot? In *In Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'07)*, 2007.

[16] Masaki Ogino, Ayako Watanabe, and Minoru Asada. Mapping from facial expression to internal state based on intuitive parenting. In *Proceedings of the Sixth International Workshop on Epigenetic Robotics*, pages 182–183, 2006.

[17] P-Y Oudeyer. Intelligent adaptive curiosity: a source of self-development. In L. Berthouze, H. Kozima, C.G. Prince, G. Sandini, G. Stojanov, G. Metta, and C. Balkenius Eds, editors, *Proc. of the Fourth Int. Workshop on Epigenetic Robotics: Modelling Cognitive Development in Robotics Systems*, volume 117, pages 127–130, 2004.

[18] P-Y Oudeyer, F. Kaplan, V.V. Hafner, and A. Whyte. The playground experiment: task-independent development of a curios robot. In D. Bank and editor Meeden, L., editors, *Proc. of the AAAI Spring Symposium Workshop on Developmental Robotics*, pages 42–47, 2005.

[19] G.J. Quirk, R.U. Muller, J.L. Kubie, and Jr. J.B. Ranck. The positional firing properties of medial entorhinal neurons: Description and comparison with hippocampal place cells. *Journal of Neuroscience*, 12(5):1945–1963, 1992.

[20] J. Schmidhuber. Curious model-building control systems. In *Proc. Int. Conf. on Neural Networks*, volume 2, pages 1458–1463. IEEE Press, 1991.

[21] M. G. Shuler and M. F. Bear. Reward timing in the primary visual cortex. *Science*, 17-311(5767):606 – 1609, 2006.

[22] J. Tani and S. Nolfi. Learning to perceive the world as articulated: An approach for hierarchical learning in sensory-motor systems. *Neural Networks*, 12(7-8):1131–1141, 1999.

[23] E.C. Tolman. Cognitive maps in rats and men. *The Psychological Review*, 55(4):189–208, 1948.