Trying anyways: how ignoring the errors may help in learning new skills

Beata J. Grzyb*†, Joschka Boedecker†, Minoru Asada†, Angel P. del Pobil* and Linda B. Smith‡
*Robotic Intelligence Laboratory, Jaume I University
Castellon de la Plana, 12071, Spain
Email: {grzyb,pobil}@icc.uji.es
† JST ERATO Asada Project, Graduate School of Eng., Osaka University
2-1 Yamadaoka, Suita, Osaka, 565-0871 Japan
Email: {joschka.boedecker,asada}@ams.eng.osaka-u.ac.jp
‡ Cognitive Development Lab, Indiana University
1101 East Tenth Street Bloomington, IN 47405-7007
Email: smith4@indiana.edu

***Abstract*—Traditional view stresses the role of errors in the learning process. The result obtained from our experiment with older infants suggested that omitting the errors during learning can also be beneficial. We propose that a temporal decrease in learning from negative feedback could be an efficient mechanism behind infant learning new skills. Herein, we claim that disregarding the errors is tightly connected to the sense of control, and results from extremely high level of self-efficacy (overconfidence). Our preliminary results with a robot simulator serve as a proof-of-concept for our approach, and suggest a possible new route for constraints balancing exploration and exploitation in intrinsically motivated reinforcement learning.**

I. Introduction

Infants do not wait for the change to come, they struggle hard to make change happen. At the beginning nothing seems to be reachable for them, as they do not have enough skills to coordinate eye and hand movements. That, however, does not stop them from trying when new interesting objects are placed in front of them. Soon these first uncoordinated movements become successful, and infants start indulging themselves in free play with surrounding objects. When the near space has been explored, again unreachable, but interesting far objects encourage young infants to try new strategies to obtain them. They can scream hoping that caregivers would understand their intentions and give the desired object to them, but nothing seems to be better and more rewarding than getting things on their own.

The main topic of the intrinsically motivated approach to reinforcement learning is to find out what causes agents to constantly increase their capabilities by exploring the world. The idea of designing models of intrinsic motivation for artificial learning systems is not new, and a significant number of models driven by novelty [1][2][3], curiousity [4], and based on competence [5][6][7] have been proposed. Oudeyer and Kaplan provided an extensive review of much of literature on intrinsic motivation systems for artificial agents [8], and proposed a typology of computational approaches indicating many possible directions in this relatively new field [9].

The tradeoff between exploration and exploitation was studied extensively in cognitive developmental robotics [10], as well as in the theory of computational reinforcement learning (CRL). In CRL the goal is to maximize the global reward, therefore the agent needs to rely on actions that led to high rewards in the past. However, if the agent is too greedy and neglects exploration it might never find the optimal strategy for the task. Infants during their development face similar problems. In order to find the best ways to perform an action they need to find a balance between exploration and exploitation. The task-independent mechanisms that could regulate this balance are not well understood [11].

We share the view of Oudeyer and Kaplan [9] that competence-based models have large potential for future research, which has not yet been explored. However, we think that these models should be more grounded in psychological experiments. The result obtained from our experiment with older infants sheds light on a possible mechanism behind infants' learning new skills [12][13]. Infants during the transition phase to walking showed a decreased ability to learn what lies within their reachable space. We hypothesize that the blocked ability to learn from negative outcome while reaching makes infants fine-tune their walking skill. Moreover, we suggest that omitting the errors may be tightly related to infants' perception of the sense of control. Our approach is based on the assumption that infants are causal agents that have the innate need of having control over the environment [14]. It has been demonstrated that people's ability to gain and maintain a sense of control is essential for their evolutionary survival [15]. The discrepancy between the actual and desired sense of control that results in frustration could contribute to more explorative behavior and discovery of walking. On the other hand, overconfidence that comes after overcoming prolonged frustration may lead to ignoring the negative feedback and contribute to more exploitative behavior. In this paper, we propose that sense of control could be a possible mechanism for balancing exploration and exploitation while learning new skills.

This paper is organized as follows. The next section introduces basic concepts of our experiment with older infants along with a short discussion on the main finding from this work. In section III, we present some neuroscientific examples where decreased learning from negative feedback was observed. In section IV we attempt to search for neuroscientific bases of our experimetal finding. Section V introduces basic concepts of our approach and provides the details of our experiment with a simulated robot. We close the paper with conclusions and discussions of follow-up research.

II. Observation data

The principal motive for our experiment was to see how infants' knowledge about their own body capabilities changes with the acquisition of new actions. A reaching action was a good candidate for our test, as to sucessfully perform this action infants need to know not only the distance to the object, but also how far they can reach and lean forward without losing balance. A total of 16 infants constituted the sample. Half were 9-month-olds (mean age, 9 months and 9 days; range, 8 months and 14 days to 9 months and 21 days), and half 12-month-olds (mean age, 11 months and 17 days; range, 11 months and 1 day to 12 months and 9 days). The basic setup of the experiment is shown in Fig. 1. The procedure of the experiment was like the following (for the details please refer to [12]). Participants were seated in a specially adapted car seat with the seatbelts fastened for security reasons. In order to keep infants engaged and attentive during the entire experimental session, a colorful stimuli display was placed in front of them. The colorful display also helped in separating the experimenter from the infants, making communication between infants and the experimenter impossible. A ball attached to a wooden dowel appeared through the opening of the frame at various distances (30, 37, 47, 60, 70cm). The sequence of trials consisted of 9 distances and begun and ended with trials at close distances to maintain infants motivated. The order of distances, apart from the first two and the last two trials in the sequence was chosen pseudo-randomly. The sequence of distances was repeated up to three times. There was no explicit reward provided to the infants after the trial for any tested distance. This helped us to avoid situations where infants could learn to make reaching movements just to communicate their interest in obtaining a reward. The entire experimental session was recorded with two cameras. These recordings were subsequentially viewed and infants' behavor scored.

The results of the experiments showed that 12-month-old, but not 9-month-old infants constantly reached for the out-of-reach objects, which was quite surprising as typically we would expect older infants to know more than younger ones. As 12 months is the age around when the transition to walking occurs, we decided to extend our experiment and recruit more infants depending on their walking abilities [13]. A sample constituted of 24 infants (mean age, 11 months and 25 days; range, 11 months and 1 day to 12 months and 22 days) categorized into 3 equal number groups, that is non-walkers, walkers with help, walkers without help. To see how reaching for far objects changes during the experimental session, we calculated the mean percentage of reaches for far distances for every sequence of trials. Fig. 2 shows the results for 60 cm distance.

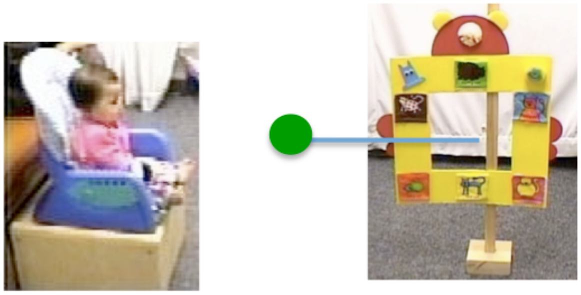


Fig. 1. Experimental setup.

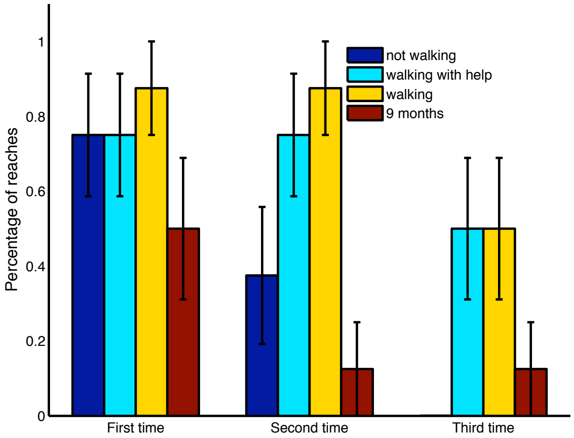


Fig. 2. Mean percentage of reaches to far objects (60cm) for 12-month-old infants: not able to walk (navy blue), able to walk with help (light blue), or able to walk without help (yellow), and 9-month-old infants (red). Please notice that we use the term "time" here, and that these are not consecutive trials. There are several trials to various distances between the first and the second time presented for a given distance.

All 12-month-old infants reached for the out-of-reach object the first time, but only walkers (with or without help) continued reaching the second and the third time. Herein, the first question arises: why 12-month-old infants reach for the far objects at all? One of the possible explanations could be the heightened interest in their surroundings. According to Zelazo [16] the attentional shifts to far objects may be caused by the growing capacity to generate functional associations in 12-month-old infants. He claimed that the qualitative change in cognitive ability gives objects new meaning and provides additional motivation for locomotion to occur by piquing the infants' interest in distal events, and thus, stimulating the use of erect locomotion. Our results, indeed, showed that 12-month-old infants reached significantly more, for the first time, for the far objects than 9-month-old infants, which could support Zelazo's position. The second question, however, is why walkers continued to reach regardless of the outcome? We suggest that a major developmental constraint for infants to learn to walk is ignoring the outcome of reaching actions.

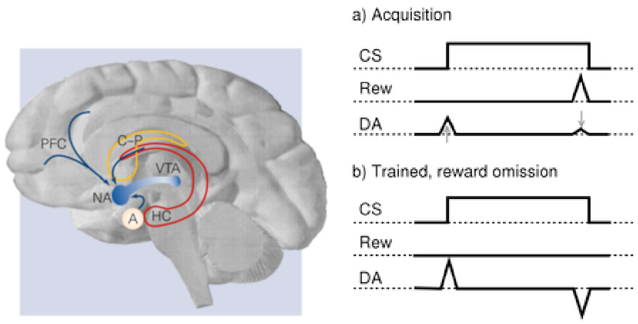


Fig. 3. On the left: the neurocircuitry of a reward system (from [18]). On the right: reward prediction error response of single dopamine neuron (from [17]).

III. Ignoring the errors

This section introduces examples of brain mechanisms that might lead to omission of errors during feedback processing, and therefore be helpful in explaining the main finding from our reaching study with older infants. Of particular interest here are dopamine neurons in the Basal Ganglia that appear to mimic the error function between the estimated reward and the actual reward received [17]. The schematic picture of this process is shown in Fig. 3. The dotted lines represent the firing baseline of dopamine neurons. During the acquisition process the dopamine cells (DA) increase firing rates when reward (Rew) is received but not expected. Over time this increase in firing rate is back propagated to the earliest reliable stimulus (CS) for the reward. The dopamine cells no longer increase their firing rate upon presentation of the predicted reward. However, when rewards are expected but not received, the firing of dopamine neurons drops below tonic baseline levels. Experiments with patients with Parkinson's disease shed light on the possible role of dopamine in trial-and-error learning in humans. As it is commonly known, patients with Parkinson's disease are characterized by a large deficiency in dopamine neurons. Unmedicated individuals are much better at learning from negative feedback than from positive feedback. Medication, however, reverses these biases and medicated individuals with Parkinson's disease are better at learning from positive than from negative feedback. These medication effects were nicely explained by Frank's basal ganglia model [19]. The errors in reward prediction are signaled by a decrease in the firing rate of dopamine neurons. The medications, however, reduce dopamine dips during negative prediction errors, and such blunting of negative prediction errors reduces learning from negative outcomes.

The inability to learn from negative feedback was shown in healthy subjects during the trust game [20]. In this experiment information about the moral profile of the oponent was provided to the players before the game started. This information can create a prior belief, but feedback from the game should adjust this prior belief to reflect new evidence. However, the experiment showed the lack of differential responses between the positive and negative outcomes when playing with morally good or bad partners. More specifically the activation of the caudate nucleus differentiated between positive and negative feedback, but only for the 'neutral partner', and not for the 'good' one, and only weakly for the 'bad' one. The normal trial-and-error learning would predict a sharp decrease in the feedback response following violations of expectations. One of the possible explanations suggested by the authors was that participants had a reward reaction to the presentation of the morally good partner, irrespective of decision.

In patients with bipolar disorder, failures in motor learning may result from the lack of striatal error signal during unsuccessful motor inhibition. Such deficits in motor regulation could be related to the emotional disregulation, as irritability and decreased motor inhibition may be linked mechanistically [21]. The impulsivity was suggested to represent a core characteristic of the disorder and to be responsible for symptoms like hyperactivation, excitability, and hasty decision making [22]. Patients with bipolar mania tend toward high goal setting, have unrealistically high success expectancies [23], and exhibit increased goal-directed activity and excessive involvement in pleasurable activities that have a high potential of risk [24]. Bipolar patients show elevated activation of dopaminergic brain areas when expecting high rewards compared to anticipation of no rewards, which could result from dysfunctional nucleus accumbens activation during prediction error processing [25]. When both, schizophrenia patients and healthy controls, showed lower nucleus accumbens activation upon omission rather than upon receipt of rewards as a potential correlate of such a learning signal, bipolar manic patients did not display a similar reduction in the activation of dopaminergic brain regions.

Infants' temporal decreased ability to learn from negative feedback while learning to walk may be related to the sense of control. The need for control has been claimed to be innate, and exercising control to be extremely rewarding and beneficial for an individual's wellbeing [14]. The newly walking infants are described as "euphoric" in relation to the first steps away from their mother [26], which could imply that the global level of dopamine is extremely high. Having control is positively correlated with activation of prefrontal cortex and negatively with the amygdala. The connections between the amygdala and the prefrontal cortex are bidirectional and appear to be essential in judging rewarding or aversive outcomes of actions. The simplest possible explanation for decreased learning from negative feedback is that exercising control is highly rewarding itself and even if the outcome of the action is not as predicted, still the reward for gaining control is provided. Another explanation could be as follows. The prefrontal cortex modulates the ventral striatal dopamine function. This regulation is biphasic and, under experimental conditions that may have relevance to pathophysiological states. The prefrontal cortex provokes an abnormal increase in the limbic dopamine function [27]. Prefrontal cortex stimulation at normal activity provides an inhibitory control over nucleus accumbens dopamine release, but prefrontal cortex stimulation at much higher than normal levels increases nucleus accumbens dopamine.

IV. Sense of control in balancing exploration and exploitation

The basic premise of our approach is that a need for control is innate, and exercising control is extremely rewarding and beneficial for an individual's wellbeing [14]. *Sense of control*, in our understanding, is one's subjective sense of the capacity to successfully perform a desired action, fulfill the individual personal goals and desires, or instinctual drives and needs. A lack of such an ability causes the feeling of frustration, and decreases the overall sense of control. On the other hand, the experience of overcoming a very difficult challenge after prolonged frustration due to many trials and errors may result in an increase of one's sense of control. Therefore, frustration and sense of control are inversely related. In line with Wong's suggestion [28], we assume that a medium (optimal) level of frustration leads to more explorative behavior, while low levels lead to exploitation.

A. Frustration and exploration

The timing of infants' transition to upright locomotion was associated with temperament [29]. More specifically, earlier walkers become more easily frustrated and stressed when physically constrained. They also reveal more persistence in reaching a blocked goal as compared to later walkers during the transition to walking [30]. We suggest that being easily frustrated could be caused by the perception of limits of self-efficacy. As suggested by Zelazo [16] 12-month-old infants are more skilled in making associations, and that may stimulate their interest in distant objects. The failures in obtaining these new challenging goals may significantly decrease infants' sense of control, increasing at the same time their level of frustration. In our opinion, growing emotional distress associated with a decreasing level of control in pre-walking infants can trigger the process of exploration. Fustration-motivated exploration, as proposed by Wong, may play the function of widening the scope of an agent's response reportoire [28].

Frustration can be represented as a simple leaky integrator as it captures the dynamics of a rapid rise in frustration level and also the possible rapid decrease over time if no input is provided:

$$df/dt = -L * f + A_o \tag{1}$$

where f is the current level of frustration, A_o is the outcome of the action and L is the fixed rate of the 'leak' ($L = 1$ in our simulations).

In classical reinforcement learning, one possibility for the agent to choose an action is a *softmax* action selection rule [31]:

$$P_t(a) = \frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}}; \tag{2}$$

where $P_t(a)$ is a probability of selecting an action a, $Q_t(a)$ is a value function for an action a, and τ is a positive parameter called the temperature that controls the stochasticity of a decision. We suggest that the frustration level can be used as a temperature parameter.

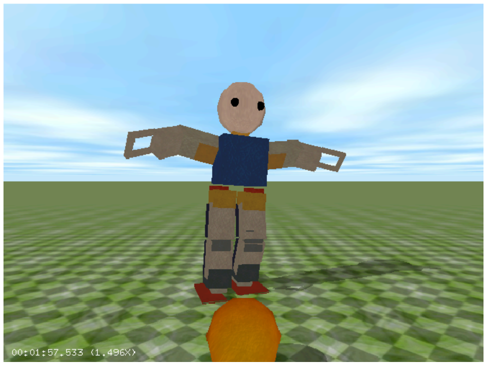
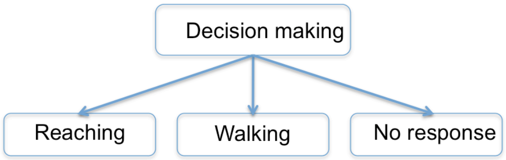


Fig. 4. The M3-neony robot simulator.


Decision making

Reaching

Walking

No response

Fig. 5. Hierarchical structure of RL modules.

B. Elation and fine-tuning

The newly walking infants are described as "euphoric" in relation to the first steps away from their mother [26]. The experience of overcoming a prolonged state of frustration that was caused by an inability to reach for a desired distant object results in an extremely high level of sense of control. We call such a state *elation*, and relate it to a sudden decrease of frustration. As the result of our experiment suggested, low learning rate may be helpful in fine-tuning the newly learned behavior. Therefore the state of elation should temporarily decrease the learning rate. In temporal difference reinforcement learning a value function V_t is updated after each choice has been made, according to the following formula:

$$V_{t+1}(c_t) = V_t(c_t) + \alpha_v * \delta_t; \tag{3}$$

where V_t is a value function, c_t a set of options, α_v is a free learning rate parameter and δ_t is the difference between the received and expected reward amounts. This formula has been adapted from [32], for a detailed description of temporal difference learning algorithm please refer to [33].

V. Simulation

We investigated how ignoring the errors could help a robot (Fig. 4) to learn new skills in an approximate optimal control framework. For the purpose of our study, the framework had a simple hierarchical structure (shown in Fig. 5). Herein, only the top most module, that is a decision making module, and the walking module were trained using simple Q-learning algorithm.

The walking module had 6 different predefined states and actions, each state was described by 8 joint angles (4 for each leg). The goal of the module was to learn how to alternate from one state to another so that the robot does not loose balance,

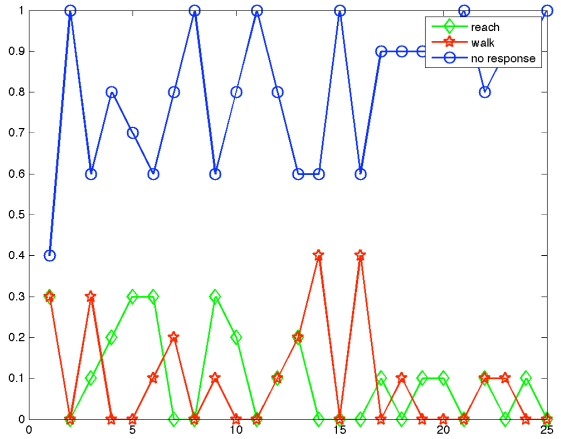


Fig. 6. Percentage of different decisions made by robot without elation.

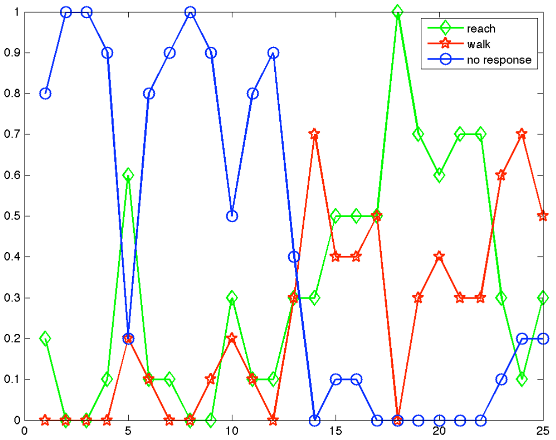


Fig. 7. Percentage of different decisions made by the robot with elation.

and it moves forward at the same time. The module received a partial reward for getting closer to the goal ($r = 10$), and negative reward for moving backwards ($r = -3$). When the robot reached the goal the module received additional reward ($r = 60$). Any action that ended up with loosing balance was punished ($r = -30$).

The state space of the decision making module was a discretized distance to the goal (6 states in our case changing by $2cm$). The goal of the modul was to select one of the possible sub-modules depending on their predicted action outcome. The module received a reward ($R = 60$) when the selected action was successful, and a punishment ($R = -30$) in the opposite case.

The simulation started with a ”young” robot, that was not able to walk. The action of walking was available for selection, but its execution did not bring any result. We simulated the onset of walking at $w = 40$ epochs. The distance to the object (close or far distance) until the onset of walking changed randomly with the probability of change 40%. After the onset of walking, the object was placed only far away. We tested the robot in two different scenarios: without state of elation, and with state of elation. The state of elation was simulated by ignoring the negative outcomes of the actions in the decision making layer.

The results of the simulations (after the ”walking onset”) are shown in Fig. 6 and Fig. 7. As it easily can be seen, the robot in first case learned that the object is not reachable, and the probability of selecting the ”no response” was very high during the entire experiment. On the other hand, the robot with elation, after 13 epochs started to select more frequently walking and reaching behavior making it possible for a walking module to improve.

VI. Discussion

In terms of the dynamic systems approach [34], we may conceptualize the role disregarding the error as follows. Assuming that the behavior of the infant is governed by a dynamic system component for decision making, and another one for execution of movement, the performance-dependent reward signal would be one of the control parameters of the decision making component. In the stable case where behaviors have been learned well (for instance to reach for near objects), negative rewards during exploratory actions would lead to further stabilization of the already learned attractors. If, however, the negative reward is ignored, i.e. the control parameter is changed and existing attractors might be destabilized. This in turn would make it easier for the system to switch to other attractors, giving their corresponding movements more chance to be practiced in a new context where they would normally not be chosen. Over time, this practice might lead to new stable attractors even under consideration of the error signal once the effect of the elated state wears off.

Although this paper focused on the motivation-based explanation for our experimental finding, other explanations are also possible. An upright posture affords greater distances for reaching, but since none of the babies had much experience with this new posture it is likely that they are not yet scaling the distances appropriate to their body size. Another explanation could be that infants may perceive far objects as *reachable by walking*. The basic assumption is that planning and coordination of walking and reaching behaviors are only possible when a certain level of the infant’s walking proficiency has been achieved, and the infant has sufficient cognitive capacity to process and store the action plan. Once action planning before the movement onset becomes possible, far objects are being represented as reachable by walking. Therefore older walkers, more frequently intend to make contact with objects placed outside of their sphere of prehension, because these objects would normally be approachable by walking. Similarly older infants may not be able to mentally immobilize the body’s remaining degrees of freedom while making decision on an object’s reachability. Standing upright increases the potentially relevant degrees of freedom and adds a balance constraint. In the sitting position infants may estimate the reachability of the object using all available degrees of freedom, not onlythe ones specific for this posture.

VII. Future work

Although our experiment suggested a possible mechanism behind infants' learning to walk, we believe that it can be extended to a more general form of intrinsically motivated open-ended learning. As the preeliminary result with the robot simulator seems to confirm the viability of our approach, the next step in our research is to perform series of experiments with a real M3-neony humanoid robot.

VIII. Conclusion

This paper presented a mechanism for balancing the exploration and exploitation while learning new skills. The core idea behind the model was that the level of sense of control determines how much the negative outcome of the action is taken into account for decision making. Omission of the errors was suggested to enable selection of different behaviors in a context when they normally would not be selected. Thereby, providing more learning opportunities for fine-tuning these behaviors. The plausibility of this mechanism was tested using a simulated humanoid robot, and our preliminary results showed strong analogy to the result obtained from our experimental data.

Acknowledgment

This research was supported by: Ministerio de Ciencia y Innovacion (FPU grant AP2007-02565); Grants-in-Aid for Scientific Research (Research Project Number: 22220002). The authors would like to thank Fabio Dalla Libera for providing the source code of the M3-neony robot simulator.

References

[1] J. Weng, "A theory for mentally developing robots," in *Proc. 2nd Int. Conf. Development Learn.*, 2002.

[2] A. Barto, S. Singh, and N. Chentanez, "Intrinsically motivated learning of hierarchical collections of skills," in *Proc. 3rd Int. Conf. Development Learn.*, 2004.

[3] J. Marshall, D. Blank, and L. Meeden, "An emergent framework for self-motivation in developmental robotics," in *Proc. 3rd Int. Conf. Development Learn.*, 2004.

[4] S. Roa, G.-J. M. Kruijff, and H. Jacobsson, "Curiosity-driven acquisition of sensorimotor concepts using memory-based active learning," in *Proceedings of the 2008 IEEE International Conference on Robotics and Biomimetics*, 2009.

[5] A. Barto and O. Simsek, "Intrinsic motivation for reinforcement learning systems," in *Proceedings of the Thirteenth Yale Workshop on Adaptive and Learning Systems*, 2005.

[6] A. Baranes and P.-Y. Oudeyer, "Maturationally-constrained competence-based intrinsically motivated learning," in *Proceedings of the Ninth IEEE International Conference on Development and Learning*, 2010.

[7] ——, "Intrinsically motivated goal exploration for active motor learning in robots: A case study," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[8] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11(2), 2007.

[9] P.-Y. Oudeyer and F. Kaplan, "What is intrinsic motivation? a typology of computational approaches," *Frontiers in Neurorobotics*, vol. 1(6), 2007.

[10] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: a survey," *IEEE Transactions on Autonomous Mental Development*, vol. 1(1), pp. 12–34, 2009.

[11] S. B. Thrun, *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches. Van Nostrand Reinhold,, NY, 1992.* New York, 1992, ch. The role of exploration in learning control.

[12] B. J. Grzyb, A. P. del Pobil, and L. B. Smith, "Reaching for the unreachable: (mis)perception of body effectivity in older infants," manuscript in preparation.

[13] ——, "Reaching for the unreachable: the cause or the consequence of learning to walk," manuscript in preparation.

[14] L. Leotti, S. Iyengar, and K. Ochsner, "Born to choose: the origins and value of the need for control," *Trends Cogn Sci.*, vol. 14(10), pp. 457–463, 2010.

[15] D. J. Shapiro, C. Schwartz, and J. Astin, "Controlling ourselves, controlling our world. psychology's role in understanding positive and negative consequences of seeking and gaining control." *Am Psychol.*, vol. 51(12), pp. 1213–30, 1996.

[16] P. Zelazo, "The development of walking: new findings and old assumptions," *J Mot Behav.*, vol. 15(2), pp. 99–137, 1983.

[17] W. Schultz, P. Dayan, and P. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275 (5306), pp. 1593–1599, 1997.

[18] [Online]. Available: http://www.cam.ac.uk/about/scienceseminars/drugs/brain.html

[19] T. V. Maia and M. J. Frank, "From reinforcement learning models to psychiatric and neurological disorders," *Nature Neuroscience*, vol. 14, pp. 154–162, 2011.

[20] M. R. Delgado, R. H. Frank, and E. A. Phelps, "Perceptions of moral character modulate the neural systems of reward during the trust game," *Nature Neuroscience*, vol. 8, pp. 1611–1618, 2005.

[21] E. Leibenluft, B. A. Rich, D. T. Vinton, E. E. Nelson, S. J. Fromm, L. H. Berghorst, P. Joshi, A. Robb, R. J. Schachar, D. P. Dickstein, E. B. McClure, and D. S. Pine, "Neural circuitry engaged during unsuccessful motor inhibition in pediatric bipolar disorder," *Am J Psychiatry*, vol. 164, pp. 52–60, 2007.

[22] P. Najt, J. Perez, M. Sanches, M. Peluso, D. Glahn, and J. Soares, "Impulsivity and bipolar disorder," *European Neuropsychopharmacology*, vol. 17, pp. 313–320, 2007.

[23] S. Johnson, "Mania and dysregulation in goal pursuit: a review," *Clin Psychol Rev*, vol. 25, pp. 241–262, 2005.

[24] "American psychiatric association (2000). diagnostic and statistical manual of mental disorders. washington, dc."

[25] B. Abler, I. Greenhouse, D. Ongur, H. Walter, and S. Heckers, "Abnormal reward system activation in mania," *Neuropsychopharmacology*, vol. 33, pp. 2217–2227, 2008.

[26] Z. Biringen, R. N. . Emde, J. J. Campos, and M. I. Appelbaum, "Affective reorganization in the infant, the mother, and the dyad: the role of upright locomotion and its timing," *Child Development*, vol. 66(2), pp. 499–514, 1995.

[27] M. E. Jackson, A. S. Frost, and B. Moghaddam, "Stimulation of prefrontal cortex at physiologically relevant frequencies inhibits dopamine release in the nucleus accumbens," *Journal of Neurochemistry*, vol. 78, pp. 920–923, 2001.

[28] P. T. Wong, "Frustration, exploration, and learning," *Canadian Psychological Review*, vol. 20(3), pp. 133–144, 1979.

[29] A. Scher, "The onset of upright locomotion and night wakings," *Perceptual and Motor Skills*, vol. 83, pp. 11–22, 1996.

[30] Z. Biringen, R. N. . Emde, J. J. Campos, and M. Appelbaum, "Development of autonomy: role of walking onset and its timing," *Perceptual and Motor Skills*, vol. 106, pp. 395–414, 2008.

[31] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[32] J. A. Beeler, N. Daw, C. R. Frazier, and X. Zhuang, "Tonic dopamine modulates exploitation of reward learning," *Frontiers in behavioral neuroscience*, vol. 4, pp. 1–14, 2010.

[33] R. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3(1), pp. 9–44, 1988.

[34] E. Thelen and L. Smith, *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, 1994.