frontiers
IN NEUROSCIENCE

ch of the neural circuits of intrinsic motivation

plan[1,*] and Pierre-Yves Oudeyer[2]

nique Federale de Lausanne, EPFL-CRAFT, Lausanne, Switzerland
er Science Laboratory Paris, Paris, France

Peter Dayan, Gatsby Computational Neuroscience Unit, University College London, UK
Kenji Doya, Neural Computation Unit, Okinawa Institute of Science and Technology, Japan

o acquire new know-how in a continuous and open-ended manner. In this paper, we hypothesize that an intrinsic motivation earning is at the origins of the remarkable structure of children's developmental trajectories. In this view, children engage nd playful activities for their own sake, not as steps toward other extrinsic goals. The central hypothesis of this paper is motivating activities correspond to expected decrease in prediction error. This motivation system pushes the infant to dictable and unpredictable situations in order to focus on the ones that are expected to maximize progress in learning. mputational model and a series of robotic experiments, we show how this principle can lead to organized sequences of easing complexity characteristic of several behavioral and developmental patterns observed in humans. We then discuss uitry underlying such an intrinsic motivation system in the brain and formulate two novel hypotheses. The first one is that acts as a learning progress signal. The second is that this progress signal is directly computed through a hierarchy of rcuits that act both as prediction and metaprediction systems.

nsic motivation, curiosity, exploration, dopamine, cortical microcircuits, meta-learning, development

CTION

nth-old toddler playing with a plastic toy car. He grasps it from different angles, puts it on the floor again, pushes akes it turn over, and sometimes by chance, manages to en, he spends some time banging the toy on the floor to g sounds but after a moment he seems to lose interest ty. As he looks around, he sees just a few steps away an ortunately left on the floor. He walks to this novel exciting dologically starts to tear it into pieces.

child suddenly lose interest in his current activity to ne? What is generating curiosity/interest/exploration in 'e will not have understood a crucial part of children's ng capabilities until we will be able to understand the that led to such kinds of organized behavior sequences. amental characteristics of children's development seem the way they explore their environment.

seem to acquire new know-how in a continuous and ner. Their capabilities for acting and perceiving contin- level of sophistication as they engage in increasingly . In just a few months, children learn to control their body, ween themselves and others, recognize sounds, smells, patterns and other multimodal situations, interact with ts, crawl, stand, walk, jump, hop, run, treat others as , participate with them in joint attention processes, in non-verbal and verbal communication, exchange shared meanings and symbolic references, play games, engage in pretend play, and eventually integrate society as autonomous social beings. A significant amount of data describes how new skills seem to build one upon another, suggesting a continuum between sensory-motor development and higher cognitive functions (Gallese and Lakoff, 2005). But the driving forces that shape this process remain largely unknown.

Second, childrens' developmental trajectories are remarkably structured (Thelen and Smith, 1994). Each new skill is acquired only when associated cognitive and morphological structures are ready. For example, children typically learn first to roll over, then to crawl, and sit, and only when these skills are operational, do they begin to learn how to stand. Likewise, sudden transitions occur from apparent insensitivity to input to stages of extraordinary sensitivity to new data. Some pieces of information are simply ignored until the child is ready for them. It is as if children were born equipped with natural means for measuring and handling complexity in order to learn in the most effective way.

Most existing views fail to account for the open-ended and self-organized nature of developmental processes. Development is either reduced to an innately defined maturational process controlled by some sort of internal clock, or, in contrast, pictured as a passive inductive process in which the child or the animal simply catches statistical regularities in the environment (see (Karmiloff-Smith, 1992; Thelen and Smith, 1994) for a critical review of current views of development). More generally, epigenetic developmental dynamics as a whole are rarely addressed as an issue as research tends to focus simply on the acquisition of particu-

er extrinsic goals. Of course, adults help by scaffolding
proposing learning opportunities, but this is just help:
s decide by themselves what they do, what they are
what their learning situations are. Far from a passive
ment has to be viewed as a fundamentally active and
ess.
rcher in psychology seem to suggest that such a kind
in the human brain and that human behavior can
otivated. However, they have postulated many differ-
at the origins of what we may call curiosity or other
oloration. The central hypothesis of this paper is that
ating activities corresponds to expected decrease in
Ve argue that children (and adults) act in order to maxi-
prediction and that this incentive shape their exploratory
iewing how concepts related to intrinsic motivation sys-
elaborated and discussed in psychology, neuroscience
ning, we present a computational model of circuits that
optimize progress in prediction. Through a series of
physical robots we show how these circuits can indeed
sequences of behavior of increasing complexity, char-
y behavioral and developmental patterns observed in
mals. We then review different hypotheses about where
erlying such an intrinsic motivation system could be
ain. In particular, we discuss the putative role of tonic
gnal of progress and formulate hypotheses about neo-
acting both as prediction and metaprediction systems.
sent a novel research program to study intrinsically moti-
volving brain imagery experiments during exploratory

MOTIVATION SYSTEMS:
OF A CONSTRUCT

nts an overview of the complex history of the concept of
n system. First, it reviews psychological models of intrin-
econd, it examines how neuroscience research, despite
ostile to this kind of construct, has nevertheless exam-
anisms linked with novelty-seeking behavior. Third, it
recent machine learning models are good candidates
gap between psychological and neuroscience models,
e instantiation of intrinsic motivation system in the form
n control architectures.

activity is characterized as intrinsically motivated when
rent reward except the activity itself (Ryan and Deci,
ek and engage in such activities for their own sake and
ead to extrinsic reward. In such cases, the person seems
nt directly from the practice of the activity. Following this
ildren playful or explorative activities can be character-
insically motivated. Also, many kinds of adult behavior
o this category: free problem-solving (solving puzzles,
ative activities (painting, singing, writing during leisure
hiking, etc. Such situations are characterized by a feel-
ontrol, concentration, enjoyment and a contraction of the
kszenthmihalyi, 1991).
of investigations concerning intrinsic motivation hap-

the general tendency to explore is never satiated and is not a consummatory response to a stressful perturbation of the organism's body. Moreover, exploration does not seem to be related to any non-nervous-system tissue deficit.

Some researchers then proposed another conceptualization. Festinger's theory of cognitive dissonance (Festinger, 1957) asserted that organisms are motivated to reduce dissonance, that is the incompatibility between internal cognitive structures and the situations currently perceived. Fifteen years later a related view was articulated by Kagan stating that a primary motivation for humans is the reduction of uncertainty in the sense of the 'incompatibility between (two or more) cognitive structures, between cognitive structure and experience, or between structures and behavior'' (Kagan, 1972). However, these theories were criticized on the basis that much human behavior is also intended to *increase* uncertainty, and not only to reduce it. Human seem to look for some forms of optimality between completely uncertain and completely certain situations.

In 1965, Hunt developed the idea that children and adult look for optimal incongruity (Hunt, 1965) He regarded children as information-processing systems and stated that interesting stimuli were those where there was a discrepancy between the perceived and standard levels of the stimuli. For, Dember and Earl (1957) the incongruity or discrepancy in intrinsically-motivated behaviors was between a person's expectations and the properties of the stimulus. Berlyne (1960) developed similar notions as he observed that the most rewarding situations were those with an intermediate level of novelty, between already familiar and completely new situations. Whereas most of these researchers focused on the notion of optimal incongruity at the level of psychological processes, a parallel trend investigated the notion of optimal arousal at the physiological level (Hebb, 1955). As over-stimulation and under-stimulation situations induce fear (e.g., dark rooms, noisy rooms), people seem to be motivated to maintain an optimal level of arousal. A complete understanding of intrinsic motivation should certainly include both psychological and physiological levels.

Eventually, a last group of researchers preferred the concept of challenge to the notion of optimal incongruity. These researchers stated that what was driving human behavior was a motivation for effectance (White, 1959), personal causation (De Charms, 1968), competence, and self-determination (Deci and Ryan, 1985).

In the recent years, the concept of intrinsic motivation has been less present in mainstream psychology but flourished in social psychology and the study of practices in applied settings, in particular in professional and educational contexts. Based on studies suggesting that extrinsic rewards (money, high grades, prizes) actually destroy intrinsic motivation (an idea actually articulated by Bruner in the 1960s (Bruner, 1962)), some employers and teachers have started to design effective incentive systems based on intrinsic motivation. However, this view is currently at the heart of many controversies (Cameron and Pierce, 2002).

In summary, most psychological approaches of intrinsic motivation postulate that ''stimuli worth investigating'' are characterized by a particular relationship (incompatibility, discrepancy, uncertainly, or in contrast, predictability) between an internal predictive model and the actual structure of the stimulus. This invites us to consider intrinsically motivating activities not only at the descriptive behavioral level (no apparent reward except the activity itself) but also primarily in respect to particular internal models built by an agent during its own personal history of interaction. To progress in the elucidation of this relationship and investigate among all the psychological models presented which are the ones really susceptible

g dynamics in brain systems are still commonly studied
xternal reward seeking (food, sex, etc.) and very rarely as
logenous and spontaneous processes. Actually, the term
en misleading as it is used in a diffent manner in neu-
in machine learning (Oudeyer and Kaplan, 2007; White,
). In behavioral neuropsychology, rewards are primarily
s or events that increased the probability and intensity
ons leading to such objects: ''rewards make you come
Thorndike, 1911). This means the function of rewards is
n behavioral effects interpreted in a specific theoretical
ultz puts it ''the exploration of neural reward mecha-
be based primarily on the physics and the chemistry
but on specific behavioral theories that define reward
z, 2006) p. 91)
nal reinforcement learning, a reward is only a numerical
rive an action-selection algorithm so that the expected
of this quantity is maximal in the future. In such con-
be thought primarily as internal measures rather than
as clearly argued by Sutton and Barto (1998)). This may
much easier from a machine learning perspective to
trinsic motivation construct as a natural extension of
learning paradigm, whereas dominant behavioral the-
mental methodology in neuroscience does not permit
construct. This is certainly one reason why com-
at do not involve any consummatory reward are rarely

e of experimental studies concerning intrinsically moti-
we can consider what resembles the most: exploratory
tended lateral hypothalamic corridor, running from the
area to the nucleus accumbens, has been recognized
of a system responsible for exploration. Pankseep calls
stem (Panksepp, 1998) (different terms are also used as
vioral activation system (Gray, 1990) or behavioral facil-
epue and Iacono, 1989)). ''This harmoniously operating
ystem drives and energizes many mental complexities
erience as persistent feelings of interest, curiosity, sen-
d, in the presence of a sufficiently complex cortex, the
meaning.'' ((Panksepp, 1998) p.145). This system, a tiny
the total brain mass, is where one of the major dopamine
(for a discussion of anatomical issue one can refer, for
, 1999; Stellar, 1985)).
functions of dopamine are known to be multiple and
ne is thought to influence behavior and learning through
ecoupled, forms of signal: phasic (bursting and pausing)
ic levels (Grace, 1991). A set of experimental evidence
mine activity can result from a large number of arous-
ing novel stimuli and unexpected rewards (Hooks and
ultz, 1998; Fiorillo, 2004). On the other hand, dopamine
ssed by events that are associated with reduced arousal
cipatory excitement, including the actual consumption
nd the omission of expected reward (Schultz, 1998).
opamine circuits appear to have a major effect on our
ment, excitement, creativity, our willingness to explore
make sense of contingencies (Panksepp, 1998). More
g evidence currently supports the view of dopamine as
of incentive salience (''wanting processes'') different
vation processes (''liking processes'') (Berridge, 2007).
A in the ventral tegmental area and of a dopamine

1973). When the dopamine system is artificially activated via electrical or chemical means, humans and animals engage in eager exploration of their environment and display signs of interest and curiosity (Panksepp, 1998). Likewise, the addictive effects of cocaine, amphetamine, opioids, ethanol, nicotine and canabinoid are directly related to the way they activate dopamine systems (Carboni et al., 1989; Pettit and Justice, 1989; Yoshimoto et al., 1991). Finally, too much dopamine activity are thought to be at the origins of uncontrolled speech and movement (Tourette's syndrome), obsessive-compulsive disorder, euphoria, overexcitement, mania and psychosis in the context of schizophrenic behavior (Bell, 1973; Grace, 1991; Weinberger, 1987; Weiner and Joel, 2002).

Things get even more complex and controversial when one tries to link these observation with precise computational models. Hypotheses concerning phasic dopamine's potential role in learning have flourished in the last ten years. Schultz and colleagues have conducted a series of recording of midbrain dopamine neurons firing patterns in awake monkeys under various behavioral conditions which suggested that dopamine neurons fire in response to unpredicted reward (see Schultz, 1998 for a review). Based on these observations, they develop the hypothesis that phasic dopamine responses drive learning by signalling an error that labels some events as ''better than expected''. This type of signalling has been interpreted in the framework of computational reinforcement learning as analogous to the prediction error signal of the temporal difference (TD) learning algorithm (Sutton, 1988). In this scheme, a phasic dopamine signal interpreted as TD-error plays a double role (Baldassarre, 2002; Barto, 1995; Doya, 2002; Houk et al., 1995; Khamassi et al., 2005; Montague et al., 1996; Schultz et al., 1997; Suri and Schultz, 2001). First, this error is used as a classical training signal to improve future prediction. Second, it is used for finding the actions that maximize reward. This so-called actor-critic reinforcement learning architecture have been presented as a relevant model to account for both functional and anatomical subdivisions in the midbrain dopamine system. However, most of the simple mappings that were first suggested, in particular the association of the actor to matrisome and the critic to the striosome part of the striatum are now seriously argued to be inconsistent with known anatomy of these nuclei (Joel et al., 2002).

Computational models of phasic dopamine activity based on the error signal hypothesis have also raised controversy for other reasons. One of them, central to our discussion, is that several stimuli that are *not* associated with reward prediction are known to activate the dopamine system in various manner. This is in particular the case for novel, unexpected ''never-rewarded'' stimuli (Fiorillo, 2004; Hooks and Kalivas, 1994; Horvitz, 2000, 2002; Ikemoto and Panksepp, 1999). The classic TD-error model does account for novelty responses. As a consequence, Kakade and Dayan suggested to extend the framework including for instance ''novelty bonuses'' (Kakade and Dayan, 2002) that distort the structure of the reward to include novelty effects (in a similar manner that 'èxploration bonuses'' permit to ensure continued exploration in theoretical machine learning models (Dayan and Sejnowski, 1996)). More recently, Smith et al. (2006) presented another TD-error model model in which phasic dopamine activation is modeled by the combination of ''Surprise'' and ''Significance'' measures. These attempts to reintegrate novelty and surprise components into a model elaborated in a framework based on extrinsic reward seeking may successfully account for a larger number of experimental observations. However, this is done in the expense of a complexification of a model that was not meant to deal with such type of behavior.

Some authors developed an alternative hypothesis to the reward prediction error interpretation, namely that dopamine promotes behavioral switching (Oades, 1985; Redgrave et al., 1999). In this interpretation

salience hypotheses, despite their psychological foun-
et supported by many computational models. But they
s in this direction. In 2003, McClure et al. (2003) argued
ience interpretation is not incompatible with the error
and presented a model where incentive salience is
pected future reward. Another recent interesting inves-
ound in (Niv et al., 2006) concerning an interpretation
es. In this model, tonic levels of dopamine is mod-
'àverage rate of reward'' and used to drive response
aster responding) into a reinforcement learning frame-
ual model, the authors claim that their theory ''dovetails
computational theories which suggest that the phasic
hine neurons reports appetitive prediction errors and
ories about dopamine's role in energizing responses''

espite many controversies, converging evidence seems
1) dopamine plays a crucial role in exploratory and
avior, (2) the meso-accumbens dopamine system is
n component to rapidly orient attentional resources to
reover, current hypotheses may favor a dual interpre-
ne's functions where phasic dopamine is linked with
nd tonic dopamine involved in processes of energizing

g

euroscience literature, we have already discussed some
hine learning models that have lead to interesting new
neurophysiologic data. Unfortunately (but not unsur-
recent research in this field are not well known by
neuroscientists. During the last 10 years, the machine
ity has begun to investigate architectures that permit
active learning (see for instance Thrun and Pratt (1998)
t al. (1997), Cohn et al. (1996)). Interestingly, the mecha-
n these papers have strong similarities with mechanisms
field of statistics, where it is called 'òptimal experiment
1972). Active learners (or machines that perform optimal
machines that ask, search and select specific training
to learn efficiently.

ally, a few researchers have started to address the
ning motivation systems to drive active learning. The
obot controlled by such systems would be able to
lore its environment not to fulfill predefined tasks but
form of intrinsic motivation that pushes it to search
ere learning happens efficiently (Barto et al., 2004;
, 2002, Kaplan and Oudeyer, 2004; Marshall et al.,
t al., 2007; Oudeyer and Kaplan, 2006; Schmidhuber,
04). Technically, most of these control systems can be
lar types of reinforcement learning architectures (Sutton
, where ''rewards'' are not provided externally by the
self-generated by the machine itself. The term 'ìntrin-
reinforcement learning'' has been used in this context
).

esearch has largely ignored the history of the intrinsic
uct as it was elaborated in psychology during the last 50
mes reinvented concepts that existed several decades
different forms of optimal incongruity). Nevertheless,
d that they introduced a novel type of understanding
ally permit to bridge the gap between the psychological

of novelty, surprise, uncertainty and incongruity correspond approximately to unexpected prediction, or in other words, significant errors in prediction. Symmetrically, the concepts of competence, effectance, self-determination, and personal causation characterize situations where prediction is accurate, which means there are small errors in prediction. In an implicit or explicit manner, error in prediction is, therefore, crucial to most of these models.

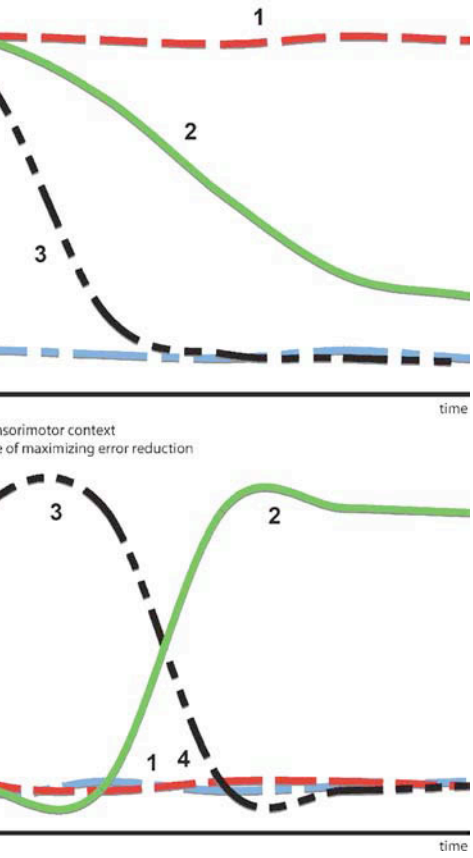
Moreover, in the neuroscience and psychological models we have discussed, the implicit idea is that the animal or person selects actions based on the prediction error. However, models differ in how this error is used. Some argue that the animal acts in order to maximize error in order to look for novel and surprising situations, others argue that it should minimize error looking for situations of mastery and a last group argue for balanced situations where incongruity is 'òptimal'' and novelty at an 'ìntermediate'' level.

Researchers conducting experiments with artificial intrinsic motivation systems have been experiencing with this design issue. From a machine learning point of view, it is relatively easy to criticize the ''maximize'' and ''minimize'' incentives. The first one pushes the animal to focus exclusively on the most unpredictable noisy parts of its environment, where learning is basically impossible. The second leads to strategies where the organism is basically immobile, avoiding novel stimulus as much as possible, which seems also a bad strategy for learning in the long term. Eventually, maintaining error at intermediary values is a too imprecise notion to permit a coherent and scalable optimization strategy.

A more interesting hypothesis would be that, in certain cases, animals and humans act in order to optimize learning progress, that is to *maximize error reduction*. This would mean that they avoid both predictable and unpredictable situations in order to focus on the ones that are expected to maximize the decrease in prediction error. In that sense, the kind of 'òptimal incongruity'' discussed in most models can be traced back to a simple principle: the search for activities where error reduction is maximal. Moreover, this model permits to articulate a direct link between a putative prediction error signal and behavioral switching patterns. **Figure 1** illustrates how a progress-driven control system operates on an idealized problem. Confronted with four sensorimotor contexts characterized by different learning profiles, the motivation for maximizing learning progress results in avoiding situations that are already predictable (context 4) or too dificult to predict (context 1), in order to focus first on the context with the fastest learning curve (context 3) and eventually, when the latter starts to reach a ''plateau,'' to switch the second most promising learning situation (context 2).

We call ''progress niches'' situations of maximal progress. Progress niches are not intrinsic properties of the environment. They result from a relationship between a particular environment, a particular embodiment and a particular time in the developmental history of the animal. Once discovered, progress niches progressively disappear as they become more predictable.

Such type of progress-driven machine learning architectures are good candidates to shed new lights on neurophysiology of intrinsic motivation. Several researchers have described models for computing learning progress. One of the first theoretical machine learning architecture implementing the principle of maximizing error reduction was described by Schmidhuber (1991), but no experiment in complex environments were conducted at that time. We have recently presented a critical discussion of the similarities and differences of these models (Oudeyer et al., 2007) and described a novel architecture capable to evaluate learning progress in complex noisy continuous environments such as the one encountered



rogress-driven control system operates on an idealized *ed with four sensorimotor contexts characterized by differ- s. The motivation for maximizing learning progress results ons already predictable (context 4) or too difficult to predict der to focus first on the context with the fastest learning and eventually, when the latter starts to reach a "plateau" econd most promising learning situation (context 2). This n system allows the creation of an organized exploratory to engage in open-ended development.*

07). In this section, we will just give a general overview of d present some robotic experiments we have performed r. Our intent is to show two things: first, that it is possible to implement an intrinsic motivation system to progress econd, that such a system permits not only to optimize to produce an organized exploration strategy and at a el to produce structured developmental patterns.

of learning progress measurement

ically motivated machine searching for learning progress ed and deep issues. The idealized problem illustrated on us to make more concrete the intuition that focusing re prediction errors decrease most can generate orga-

the world with an organized predefined set of possible kinds of activities. It would in fact be contradictory, since they are capable of open-ended development, and most of what they will learn is impossible to know in advance. It also occurs for a developmental robot, for which the world is initially a fuzzy blooming flow of unorganized sensorimotor values. In this case, how can we define learning progress? What meaning can we attribute to ''maximizing the decrease of prediction errors?''

A first possibility would be just to compute learning progress at time t as the difference between the mean prediction errors at time t and at time $t-\theta$. But implementing this on a robot quickly shows that it is in fact nonsense. For example, the behavior of a robot motivated to maximize such a progress would be typically an alternation between jumping randomly against walls and periods of complete immobility. Indeed, passing from the first behavior (highly unpredictable) to the second (highly predictable) corresponds to a large decrease in prediction errors, and so to a large internal reward. So, we see that there is a need to compute learning progress by comparing prediction errors in sensorimotor contexts that are similar, which leads us to a second possible approach.

In order to describe this second possibility, we need to introduce a few formal notations and precisions about the computational architecture that will embed intrinsic motivation. Let us denote a sensorimotor situation with the state vector $\boldsymbol{x}(t)$ (e.g., a given action performed in a given context), and its outcome with $\boldsymbol{y}(t)$ (e.g., the perceptual consequence of this action). Let us call M a prediction system trying to model this function, producing for any $\boldsymbol{x}(t)$ a prediction $\boldsymbol{y'}(t)$. Once the actual evolution $\boldsymbol{y'}(t)$ is known, the error $e_x(t) = |\ \boldsymbol{y}(t) - \boldsymbol{y'}(t)$ in prediction can be computed and used as a feedback to improve the performances of M. At this stage, no assumption is made regarding the kind of prediction system used in M. It could be for instance a linear predictor, a neural network or any other prediction method currently used in machine learning. Within this framework, it is possible to imagine a first manner to compute a meaningful measure of learning progress. Indeed, one could compute a measure of learning progress $p_x(t)$ for every single sensorimotor situation x through the monitoring of its associated prediction errors in the past, for example with the formula:

$$p_x(t) = \langle e_x(t-\theta)\rangle - \langle e_x(t)\rangle \tag{1}$$

where $\langle e_x(t)\rangle$ is the mean of $\boldsymbol{e}_x$ values in the last τ predictions. Thus, we here compare prediction errors in exactly the same situation $\boldsymbol{X}$, and so we compare only identical sensorimotor contexts. The problem is that, whereas this is an imaginable solution in small symbolic sensorimotor spaces, this is inapplicable to the real world for two reasons. The first reason is that, because the world is very large, continuous and noisy, it never happens to an organism to experience twice exactly the same sensorimotor state. There are always slight differences. A possible solution to this limit would be to introduce a distance function $\boldsymbol{d}(\boldsymbol{x}_m, \boldsymbol{x}_n)$ and to define learning progress locally in a point $\boldsymbol{x}$ as the decrease in prediction errors concerning sensorimotor contexts that are close under this distance function:

$$p_x(t) = \langle e_x^{\delta}(t-\theta)\rangle - \langle e_x^{\delta}(t)\rangle \tag{2}$$

where $\langle e_x^{\delta}(t)\rangle$ denotes the mean of all $\{e_{x_1} | d(x, x_1) < \delta\}$ values in the last τ predictions, and where δ is a small fixed threshold. Using this measure would typically allow the machine to manage to repeatedly try roughly the same action in roughly the same context and identify all the resulting prediction errors as characterizing the same sensorimotor situation (and thus overcoming the noise). Now, there is a second problem

sting only for a very brief amount of time, and will hardly oration. For example, using this approach, a robot playing might try to squash it on the ground to see the noise it ncing learning progress in the first few times it tries, but p playing with it and typically would not try to squash it e sofa or on a wall to hear the result. This is because its tial learning progress is still too local.

l measure of learning progress

le that there really is a need to build broad categories squashing plastic toys on surfaces or shooting with the cts) as those pre-given in the initial idealized problem. of learning progress will only become both meaning- if an automatic mechanism allows for the distinction es of activities, typically corresponding to not-so-small sorimotor space. We have presented a possible solution, tive splitting of the sensorimotor space into regions $\mathcal{R}_n$. rimotor space is considered as one big region, and pro- s split into sub-regions containing more homogeneous and sensorimotor contexts (the mechanisms of splitting udeyer et al., 2007]). In each region $\mathcal{R}_n$, the history of $\{e\}$ is memorized and used to compute a measure of that characterizes this region:

$$\theta)\rangle - \langle e_{\mathcal{R}}(t)\rangle \quad (3)$$

he mean of $\{e_X | X \in \mathcal{R}_n\}$ values in the last τ predictions. terative region-based operationalization of learning re two general ways of building a neural architecture plement intrinsic motivation. A first kind of architecture, includes two loosely coupled main modules. The first the neural circuitry implementing the prediction machine er, and learning to predict the $\boldsymbol{x} \rightarrow \boldsymbol{y}$ mapping. The sec- d be a neural circuitry meta M organizing the space into $\mathcal{R}_n$ and modelling the learning progress of M in each based on the inputs ($\boldsymbol{x}(t)$, $\boldsymbol{e_x}(t)$ provided by M. This es no assumption at all on the mechanisms and repre- y the learning machine M. In particular, the splitting of ions is not informed by the internal structure of M. This n of the architecture general, but makes the scalability al-world structured inhomogeneous spaces where typi- ral resources will be recruited/built for different kinds of

e have developed a second architecture, in which the meta M are tightly coupled. In this version, each region with a circuit $M_{\mathcal{R}}$, called an expert, as well as with a chine $meta\ M_{\mathcal{R}}$. A given expert $M_{\mathcal{R}}$ is responsible for $\boldsymbol{y}$ given $\boldsymbol{x}$ when $\boldsymbol{x}$ is a situation which is covered by $\mathcal{R}_n$ $M_{\mathcal{R}}$ is only trained on inputs ($\boldsymbol{x}$, $\boldsymbol{y}$) where $\boldsymbol{x}$ belongs to on $\mathcal{R}_n$. This leads to a structure in which a single expert d for each non-overlapping partition of the space. The $eta\ M_{\mathcal{R}}$ associated to each expert circuit can then com- rning progress of this region of the sensorimotor space for a symbolic illustration of this splitting/assignment a of using multiple experts has been already explored including for instance (Doya et al., 2002; Baldassarre, l Jacobs, 1994; Kawato, 1999; Khamassi et al., 2005; 99)

this can be formulated as the maximization of future expected rewards (i.e., maximization of the return), that is

$$E\{r(t+1)\} = E\left\{\sum_{t \geq t_n} \gamma^{t-t_n} r(t)\}\right\}$$

where $\gamma (0 \leq \gamma \leq 1)$ is the discount factor, which assigns less weight on the reward expected in the far future. We can note that at this stage, it is theoretically easy to combine this intrinsic reward for learning progress with the sum of other extrinsic rewards $r_e(t)$ coming from other sources, for instance in a linear manner with the formula $r(t) = \alpha . p_{\mathcal{R}}(t) + (1 - \alpha) r_e(t)$ (the parameter α measuring the relative weight between intrinsic and extrinsic rewards).

This formulation corresponds to a reinforcement learning problem (Sutton and Barto, 1998) and thus the techniques developed in this field can be used to implement an action selection mechanism which will allow the system to maximize future expected rewards efficiently (e.g., Q-learning (Walkins and Dayan, 1992), TD-learning (Sutton, 1988), etc.). However, predicting prediction error reduction is, by definition, a highly non-stationary problem (progress niches appear and disappear in time). As a consequence, traditional ''slow'' reinforcement learning techniques are not well adapted in this context. In (Oudeyer et al., 2007), we describe a very simple action-selection circuit that avoids problems related to delayed rewards and makes it possible to use a simple prediction system which can predict $r(t+1)$ and so evaluate $E\{r(t+1)\}$ Let us consider the problem of evaluating $E\{r(t+1)\}$ given a sensory context $\boldsymbol{S}(t)$ and a candidate action $\boldsymbol{M}(t)$, constituting a candidate sensorimotor context $\boldsymbol{SM}(t) = \boldsymbol{x}(t)$ covered by region $\mathcal{R}_n$. In our architecture, we approximate $E\{r(t+1)\}$ with the learning progress that was achieved in $\mathcal{R}_n$ with the acquisition of its recent exemplars, i.e. $E\{r(t+1)\} \approx p_{\mathcal{R}}(t - \theta_{\mathcal{R}})$ where $t - \theta_{\mathcal{R}}$ is the time corresponding to the last time region $\mathcal{R}_n$ and the associated expert circuit processed a new exemplar. The action–selection loop goes as follows:

- in a given sensory $\boldsymbol{S}(t)$ context, the robot makes a list of the possible values of its motor channels $\boldsymbol{M}(t)$ which it can set; If this list is infinite, which is often the case since we work in continuous sensorimotor spaces, a sample of candidate values is generated;
- each of these candidate motor vectors $\boldsymbol{M}(t)$ associated with the sensory context $\boldsymbol{S}(t)$ makes a candidate $\boldsymbol{SM}(t)$ vector for which the robot finds out the corresponding region $\mathcal{R}_n$; then the formula we just described is used to evaluate the expected learning progress $E\{r(t+1)\}$ that might be the result of executing the candidate action $\boldsymbol{M}(t)$ in the current context;
- the action for which the system expects the maximal learning progress is chosen with a probability $1 - \in$ and executed, but sometimes a random action is selected (with a probability $\in$, typically 0.35 in the following experiments).
- after the action has been executed and the consequences measured, the system is updated.

More sophisticated action-selection circuits could certainly be envisioned (see, for example, (Sutton and Barto, 1998)). However, this one revealed to be surprisingly efficient in the real-world experiments we conducted.

Experiments

We have performed a series of robotic experiments using this architecture

roduce sounds. Various toys are placed near the robot, programed 'àdult'' robot which can respond vocally to certain conditions. At the beginning of an experiment, ot know anything about the structure of its continuous ce (which actions cause which effects). Given the size of stive exploration would take a very long time and random be inefficient.

obotic experiment, which lasts approximately half a day, s of the sensorimotor channels are stored, as well as ures which help us to characterize the dynamics of the ent. The evolution of the relative frequency of the use of ators is measured: the head pan/tilt, the arm, the mouth eakers (used for vocalizing), as well as the direction in s turning its head. **Figure 3** shows data obtained during e experiment.

ng of the experiment, the robot has a short initial phase ation and body babbling. During this stage, the robot's alent to the one we would obtain using random action rly observe that in the vast majority of cases, the robot ok at or act on objects; it essentially does not interact ent.

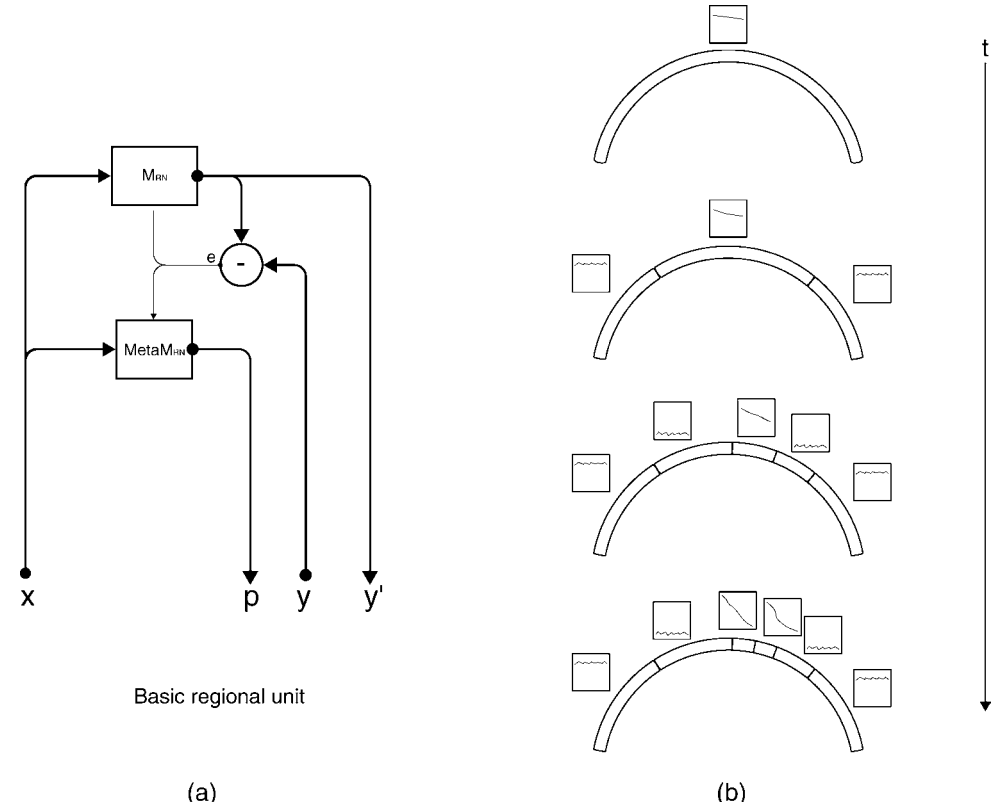
a phase during which the robot begins to focus suc- ing with individual actuators, but without knowing the lances: first there is a period where it focuses on trying ctions (and stops bashing or producing sounds), then it focuses on just looking around, then it focuses on trying to bark/vocalize toward all directions (and stops biting and bashing), then on biting, and finally on bashing in all directions (and stops biting and vocalizing).

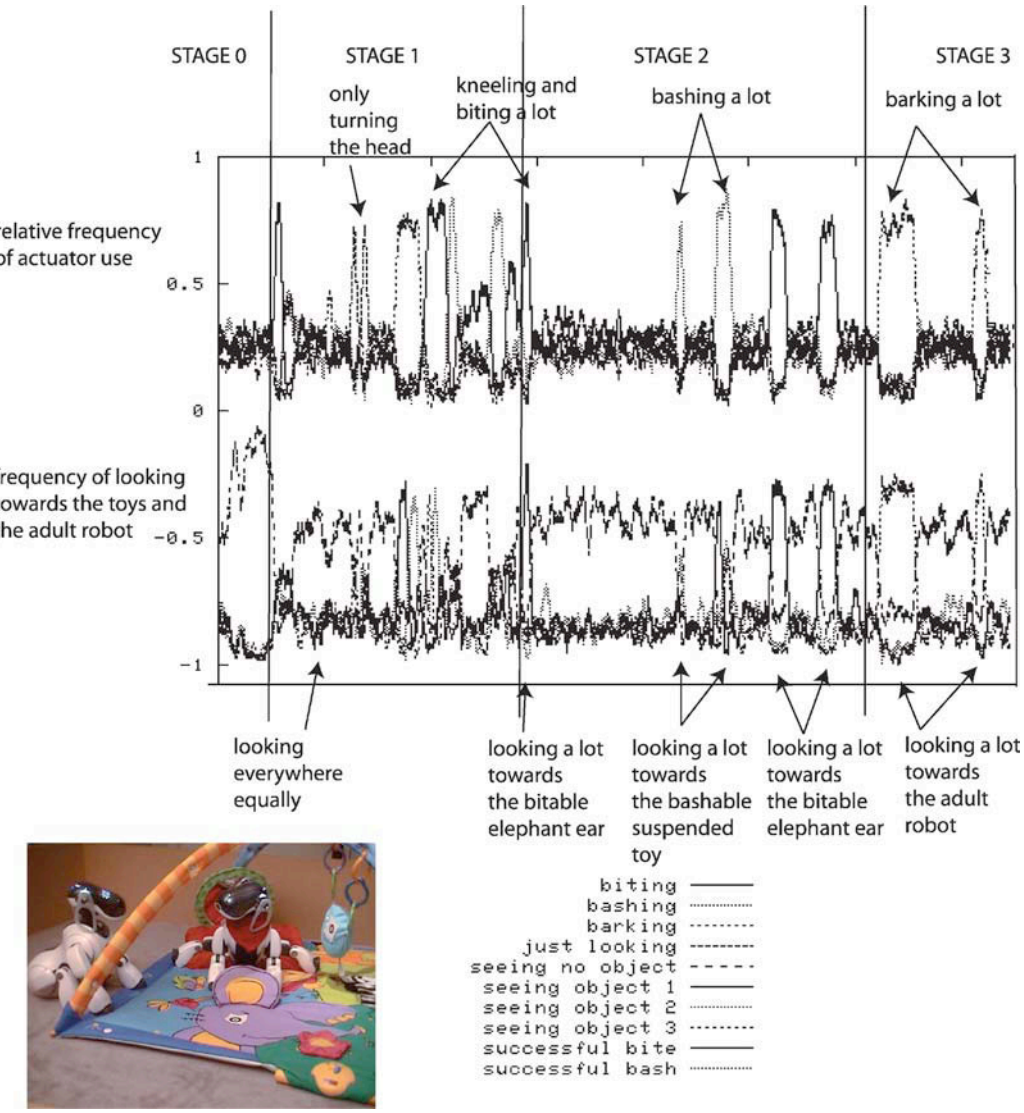
Then, the robot comes to a phase in which it discovers the precise affordances between certain action types and certain particular objects. It is at this point focusing either on trying to bite the bitable object (the elephant ear), or on trying to bash the bashable object (the suspended toy).

Eventually, it focuses on vocalizing towards the 'àdult'' robot and listens to the vocal imitations that it triggers. This interest for vocal interactions was not pre-programed, and results from exactly the same mechanism which allowed the robot to discover the affordances between certain physical actions and certain objects.

The developmental trajectories produced by these experiments can be interpreted as assimilation and accommodation phases if we retain the Piagetian's terminology (Piaget, 1952). For instance, the robot ''discovers'' the biting and bashing schema by producing repeated sequences of these kinds of behavior, but initially these actions are not systematically oriented towards the bitable or the bashable object. This stage corresponds to 'àssimilation.'' It is only later that 'àccommodation'' occurs as biting and bashing starts to be associated with their respective appropriate context of use.

Our experiments show that functional organization can emerge even in the absence of explicit internal schema structures and that developmental patterns can spontaneously self-organize, driven by the intrinsic motiva-





experimental run. *The robot, placed on a play mat, can move its arms, its neck and mouth and produce sounds. Various toys are placed near as a pre-programed "adult" robot which can respond vocally to the other robot in certain conditions. Results obtained after a typical run of the own. Top curves: relative frequency of the use of the use of different actuators (head pan/tilt, arm, mouth, sound speaker). Bottom curves: ng toward each object and in particular toward pre-programed robot.*

ave discussed elsewhere how these type of patterns are et some results from the developmental psychology and ion literature (Kaplan and Oudeyer, 2007a,b).

TIONS ABOUT THE NEURAL OF INTRINSIC MOTIVATION

nesis of this paper is that intrinsically motivating activities xpected prediction error decrease. Through a computer

Hypothesis 1: Tonic dopamine as a signal of expected prediction error decrease

We have already reviewed several elements of the current complex debate on the role and function of dopamine in action selection and learning. Based on our investigations with artificial intrinsic motivation systems, we would like to introduce yet another interpretation of the potential role of dopamine by formulating the hypothesis that tonic dopamine acts as a signal of ''progress niches,'' i.e. states where prediction error of some internal model is expected to decrease. As experimental researches in

cally motivated behavior, we believe this view is com-
ypothesis of dopamine as a signal of ''progress niches.''
s an invitation to investigate these ''promising'' states.
n is also coherent with investigations that were con-
g human affective experience during stimulation of the
s. When the lateral hypothalamus dopamine system is
of the SEEKING system previously discussed), people
at ''something very interesting and exciting is going on''
, p. 149 based on experiments reported in Heath (1963),
4)). This corresponds to subjective affective states linked
motivating activities (Csikszenthmihalyi, 1991).
erridge articulates the proposition that ''dopamine neu-
ormational consequence of learning signals, reflecting
diction that is generated elsewhere the brain but do not
earning themselves'' (Berridge (2007), p. 405]). In this
signals are a consequence and not a cause of learning
ening elsewhere in the brain. This is consistent with the
e neurons originating in the midbrain are recognized to
direct access to the signals information that needs to be
ssociative learning mechanism. All the signals that they
to be ''highly processed already by forebrain structures
cells get much learning-relevant information'' (Berridge
e also Dommett et al. (2005)).
this progress signal is used as a reinforcement to drive
nd behavioral switching. This aspect of our architecture
imilar interpretation of the role of dopamine in several
v often criticized) actor-critic models of action-selection
basal ganglia (Baldassarre, 2002; Barto, 1995; Doya,
, 1995; Khamassi et al., 2005; Montague et al., 1996;
97; Suri and Schultz, 2001). Let us recall that the dorsal
glutamate inputs from almost all regions of the cere-
al neurons fire in relation to movement of a particular
so to preparation of movement, desired outcome of a
ual and auditory stimuli and to visual saccades toward
tion. In most actor-critic computational models of the
pamine responses originating the substantia nigra is
reasing the synaptic strength, between currently active
output elements (thus shaping the policy of the actor
interpretation). With this mechanism, if the striatal out-
to motor responses and that dopamine cells become
sence of an unexpected reward, the same pattern of
it the same pattern of motor outputs in the future. One
o this interpretation is that 'if DA neurons respond to
events, regardless of appetitive or aversive values, one
hat DA activation does not serve to increase the like-
n behavioral response is repeated under similar input
itz (2002) p. 70). Progress niches can be extrinsically
rogress in playing poker sometimes result in gaining
aversive (i.e., risk-taking behavior in extreme sports).
ieve our hypothesis is compatible with interpretations of
based action-selection circuits that control the choice
cortico-striato-thalamo-cortical loops.
precise architecture of this reinforcement learning archi-
at this stage very open. A seducing hypothesis would
studied reinforcement learning architectures based on
rror phasic signals could be just reused with an internal
ward, namely expected progress. This should lead to a
terpretation of the role of phasic and tonic dopamine in

on the difference of two long-run average prediction error rate (Equation 3). We will now discuss how and where this progress signal could me measured.

Hypothesis 2: Cortical microcircuits as both prediction and metaprediction systems

Following our hypothesis that tonic dopamine acts as signal of prediction progress, we must now guess where learning progress could be computed. For this part, our hypothesis will be that cortical microcircuits act as both prediction and metaprediction systems and therefore can directly compute regional learning progress, through an unsupervised regional assignment as this is done in our computational model.

However, before considering this hypothesis let us briefly explore some alternative ones. The simpler one would be that progress is evaluated in some way or another in the limbic system itself. If indeed, as many authors suggests, phasic responses of dopamine neurons report prediction error in certain contexts, their integration over time could be easily performed just through the slow accumulation of dopamine in certain part of neural circuitry (hypothesis discussed in (Niv et al., 2006)). By comparing two running average of the phasic signals one could get an approximation of Equation 1. However, as we discussed in the previous section, to be appropriately measured, progress must be evaluated in regional manner, by local 'èxpert'' circuits. Although it is not impossible to imagine an architecture that would maintain such type of regional specialized circuitry in the basal ganglia (see for instance the multiple expert actor-critic architectures described by (Khamassi et al., 2005)), we believe this is not the most likely hypothesis.

As we argued, scalability considerations in real-world structured inhomogeneous spaces favor architectures in which neural resources can be easily recruited or built for different kinds of initially unknown activities. This still leaves many possibilities. Kawato argues that, from a computational point of view, 'it is conceivable that internal models are located in all brain regions having synaptic plasticity, provided that they receive and send out relevant information for their input and output'' (Kawato, 1999). Doya (1999) suggested broad computational distinction between the cortex, the basal ganglia, and the cerebellum, each of those associated with a particular type of learning problems, unsupervised learning, reinforcement learning and supervised learning, respectively. Another potential candidate location, the hippocampus has often been described as a comparator of predicted and actual events (Gray, 1982) and fMRI studies revealed that its activity was correlated with the occurrence of unexpected events (Ploghaus et al., 2000). Among all these possibilities, we believe the most promising direction of exploration is the cortical one, essentially because the cortex offers the type of open-ended unsupervised'èxpert circuits'' recruitment that we believe are crucial for the computation of learning progress.

A single neural microcircuit forms an immensely complicated network with multiple recurrent loops and highly heterogeneous components (Douglas and Martin, 1998; Mountcastle, 1978; Shepherd, 1988). Finding what type of computation could be performed with such a high dimensional dynamical system is a major challenge for computational neuroscience. To explore our hypothesis, we must investigate whether the computational power and evolutionary advantage of columns can be unveiled if these complex networks are considered not only as predictors but performing both prediction and metaprediction functions (by not only anticipating future sensorimotor events but also its own errors in prediction and learning progress).

eyer

ecently, Deneve et al. (2007) presented a model of a d on recurrent basis function networks, a kind of model mapped onto cortical circuits. Kalman filters share some e kind of metaprediction machinery we have discussed they also deal with modeling errors made by prediction . However, we must admit that there is not currently any nental evidence or computational model that supports that cortical circuit actually compute their own learning

could show that cortical microcircuits can signal this her parts of the brain, the mapping with our model ateral inhibition mechanisms, specialization dynamics ganizing processes that are typical of cortical plastic- without problems to perform the type of regionalization tor space that our architecture features. As previously lection could then be realized by some form of subcorti- chitecture, similar to the one involved in the optimization of rewards.

lieve our hypothesis is consistent from an evolutionary t least that an 'èvolutionary story'' can be articulated atively ''recent'' invention of the cortical column circuits with the fact that only mammals seems to display intrin- behavior. Once discovered by evolution, cortical columns nemselves leading to the highly expanded human cortex f cortical neurons (10^{10}) among all animals, closely fol- aceans and elephants (Roth and Dicke, 2005), over 1000 om mouse to man to provide 80% of the human brain). em so advantageous from an evolutionary point of view? suppose that intrinsic motivation systems appeared after isting machinery dedicated to the optimization of extrin- r an extrinsically motivated animal, value is linked with articular visual patterns, movement, loud sounds, or any that signal that basic homeostatic physiological needs cal integrity are (not) being fulfilled. These animals can l strategies to experience the corresponding situations le. However, when an efficient strategy is found, noth- further toward new activities. Their development stops

n of a basic cortical circuit that could not only acts as as metapredictor capable of evaluating its own learn- be seen as a major evolutionary transition. The brain produce its own reward, a progress signal, internal to us system with no significant biological effects on non- issues. This is the basis of an adaptive internal value sensorimotor experiences that produce positive value This is what drives the acquisition of novel skills, with re and complexity. This is a revolution, yet it is essen- e old brain circuitry that evolved for the optimization of needs. If we follow our hypothesis, the unique human n has to be understood as a coevolutionary dynamical rger ''space'' for learning and more things to learn. In uman culture, as a huge reservoir of progress niches, sure in having more of these basic processing units.

TIVES

t it is always hazardous to make too simple mappings

from an neuroscience point of view, which means getting data of what is going on in the brain during such type of exploratory behavior. In addition, we must find a way of comparing the experiments conducted with human subjects with the behavior of artificial models, which in such a context is not an easy problem.

For the adult and infant studies, experiments could consist in observing infant and adults' behavior during their exploration of virtual environment, monitoring in real-time their neural dynamics using brain imagery techniques (for instance using a similar experimental set up than the one used in (Koepp et al., 1998)). Conducting experiments in virtual world offers a number of interesting advantages compared to experiments in real physical environment. In virtual worlds, learning opportunities can be easily controlled and designed. One could for instance create a virtual environment designed so that the degree of learning opportunities becomes an experimental variable, permitting easy shifts from rich and stimulating environments to boring and predictive worlds. Moreover, virtual worlds can be made suffciently simple, abstract and novel in order to feature learning opportunities that do not depend too much on previously acquired skills. Embodiment plays a crucial role in shaping developmental trajectories and sequences of skills acquisition (see also Lakoff and Johnson (1998) in that respect). Paradoxically, this means that in order to identify novel exploration patterns and compare human trajectories with the ones of an artificial agent, we must create situations where the embodiment is radically different from natural human embodiment. In such a case, the human and the artificial agent would have to master an *(equally) unknown body in an (equally) unknown world.* To some extent this proposed approach is related to Galantucci's experiments on dynamics of convention formation in an novel, previously unknown medium (Galantucci, 2005). The expected outcome of this kind of experiment is to obtain a first characterization of the types of cortical neural assemblies involved in intrinsically motivated behavior, a kind of data which currently lacks to progress.

Research in psychology and neuroscience provides important elements supporting the existence of intrinsic motivation systems. Computational models permit to investigate possible circuits necessary for different elements of an intrinsic motivation system and to explore their structuring effect on behavioral and developmental patterns. It is likely that many diverse lines of experimental data can potentially be explained in common terms if we consider children as active seekers of progress niches, who learn how to focus on what is learnable in the situations they encounter and on what can be effciently grasped at a given stage of their cognitive and physiological development. But to progress in such an understanding, we need to define a novel research program combining infant studies, analysis of realistic computational model and experiments with robots and virtual agents. We believe that if such a research agenda can be conducted, we are about to reach a stage where it will be for the first time possible to study the cascading consequences in development that small changes in motivation systems can provoke.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential confiict of interest.

REFERENCES

Baldassarre, G. (2002). A modular neural-network model of the basal ganglia's role in learning and selection motor behaviors. *J. Cogn. Sys. Res.* 3, 5–13.

Barto, A. (1995). Adaptive critics and the basal ganglia. In *Models of information processing in the basal ganglia*. J. Houk, J. Davis, and D. Beiser, eds. Cambridge, MA

mayer, W., Hornykiewicz, Jellinger, K., and Seitelberger, F. (1973). and the syndromes of parkinson and huntington: Clinical, morphochemical correlations. *J. Neurol. Sci.* 20, 415–455.
. The debate over dopamine's role in reward: the case of incentive *pharmacology* 191, 391–431.
n Knowing: Essays for the Left Hand (Cambridge, MA, Harvard
erce, W. (2002). *Rewards and Intrinsic Motivation: Resolving the* in and Garvey Press).
, A., Perezzani, L., and Di Chiara, G. (1989). Amphetamine, cocaine, d nomifensine increases extra-cellullar dopamine concentrations the nucleus accumbens of freely moving rats. *Neuroscience* 28,
i, Z., and Jordan, M. (1996). Active learning with statistical models. s. 4, 129–145.
The Discovery of the Artifical. Behavior, Mind and Machines, Before ernetics (Kluwer academic publishers, Dordrecht).
. (1991). *Flow-the Psychology of Optimal Experience* (Harper Perenand Dayan, P. (2002). Opponent interactions between serotonin and *l Netw.* 15, 603–616.
owski, T. J. (1996). Exploration bonuses and dual control. *Mach.*
. Personal Causation: *The Internal Affective Determinants of Behav*ademic Press).
R. (1985). I*ntrinsic Motivation and Self-Determination in Human* n Press.
arl, R. W. (1957). Analysis of exploratory, manipulatory and curiosity *ol. Rev.* 64, 91–96.
, J. -R., and Pouget, A. (2007). Optimal sensorimotor integration in networks: A neural implementation of kalman filters. *J. Neurosci.*
o, W. (1989). Neurobehavioral aspects of affective disorders. *Ann.* 457–492.
V., Blatha, C., Martindale, J., Lefebre, V., Walton, N., Mayhew, nd Redgrave, P. (2005). How visual stimuli activate dopaminergic latency. *Science*, 307, 1476–1479.
tin, K. (1998). Neocortex. In *The Synaptic Organization of the Brain*, d., (Oxford University Press) PP. 459–509.
are the computations of cerebellum, basal ganglia, and the cerebral *etw.* 12, 961–974.
alearning and neuromodulation. *Neural Netw.* 15, 4–5.
K., Katagiri, K., and Kawato, M. (2002). Multiple modelbased arning. *Neural Comput.* 14, 1347–1369.
heory of Optimal Experiment (New York, NY, Academic Press).
A theory of Cognitive Dissonance (Evanston, Row, Peterson).
The uncertain nature of dopamine. *Mol. Psychiatry*, 122–123.
An experimental study of the emergence of human communication *Sci.* 29, 737–767.
f, G. (2005). The brainıs concepts: role of sensorymotor conceptual *. Neuropsychol.* 21.
Phasic versus tonic dopamine release and the modulation of n responsivity: a hypothesis for the etiology of schizophrenia. *Neu*24.
Neuropsychology of Anxiety: An Enquiry into the Functions of the pal System (Oxford, Clarendon Press).
systems that mediate both emotion and cognition. *Cogn. Emot.* 4,
arning and satiation of response in intrinsically motivated complex nces by monkeys. *J. Comp. Physiolo. Psychology* 43, 289–294.
lectrical self-stimulation of the brain in man. *American J. Psychiatry*
rives and the c.n.s (conceptual nervous system). *Psycholo. Rev.* 62,
vas, P. (1994). Involvement of dopamine and excitatory amino n in novelty-induced motor activity. *J. Pharmacol Exp. Ther.* 269,
Mesolimbocortical and nigrostriatal dopamine responses to salient nts. *Neuroscience* 96, 651–656.
). Dopamine gating of glutamatergic sensorimotor and incentive t signals to the striatum. *Behav. Brain Res.* 137, 65–74.
nd Barto, A. (1995). A model of how the basal ganglia generate and ls that predict reinforcement. In *Models of Information Processing glia*, J. Houk, J. Davis and D. Beiser eds. (MIT press) PP. 249–270.
Ikemoto, S., and Panksepp, J. (1999). The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to rewardseeking. *Brain Res. Rev.* 31, 6–41.
Jaeger, H. (2001). The echo state approach to analyzing and training recurrent neural networks. Technical Report, GMD Report 148, GMD - German National Research Institute for Computer Science.
Jaeger, H., and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science* 304, 78–80.
Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Netw.* 15, 535–547.
Jordan, M., and Jacobs, R. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural Comput.* 6, 181–214.
Kagan, J. (1972). Motives and development. *J. Pers. Soc. Psychol.* 22, 51–66.
Kakade, S., and Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Netw.* 15, 549–559.
Kaplan, F., and Oudeyer, P. -Y. (2004). Maximizing learning progress: an internal reward system for development. In Embodied Artificial Intelligence, F. Iida, R. Pfeifer, L. Steels and Y. Kuniyoshi eds. LNCS 3139, (Springer-Verlag, London, UK) PP. 259–270.
Kaplan, F., and Oudeyer, P.-Y. (2007a). The progress-drive hypothesis: an interpretation of early imitation. In *Models and Mechanisms of Imitation and Social Learning: Behavioural, Social and Communication Dimensions*, C. Nehaniv and K. Dautenhahn eds. (Cambridge University Press) PP. 361–377.
Kaplan, F., and Oudeyer, P. -Y. (2007b). Un robot motivé pour apprendre: le role des motivations intrinseques dans le developpement sensorimoteur. *Enfance* 59, 46–58.
Karmiloff-Smith, A. (1992). Beyond Modularity: A Developmental Perspective on Cognitive Science (MIT Press).
Kawato, M. (1999). Internal models for motor control and trajectory planning. *Curr. Opin. Neurobiol.* 9, 718–727.
Khamassi, M., Lachèze, L., Girard, B., Berthoz, A., and Guillot, A. (2005). Actor-critic models of reinforcement learning in the basal ganglia. *Adapt. Behav.* 13, 131–148.
Koepp, M., Gunn, R., Cunningham, V., Dagher, A., T., J., Brooks, D. J., Bench, C. J., and Grasby, P. M. (1998). Evidence for striatal dopamine release during a video game. *Nature* 393, 266–267.
Lakoff, G., and Johnson, M. (1998). *Philosophy in the Flesh: the Embodied Mind and its Challenge to Western Thought* (Basic Books).
Maas, W., Natschlager, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.* 14, 2531–2560.
Marshall, J., Blank, D., and Meeden, L. (2004). An emergent framework for self-motivation in developmental robotics. In *Proceedings of the 3rd International Conference on Development and Learning (ICDL 2004)*, Salk Institute, San Diego.
McClure, S., Daw, N., and Montague, P. (2003). A computational substate for incentive salience. *Trends Neurosci.* 26.
Montague, P., Dayan, P., and Sejnowski, T. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J. Neurosci.* 16, 1936–1947.
Montgomery, K. (1954). The role of exploratory drive in learning. *J. Comp. Physiol. Psychol.* 47, 60–64.
Mountcastle, V. (1978). An organizing principle for cerebral function: The unit model and the distributed system. In *The Mindful Brain*. G. Edelman and V. Mountcastle eds. (MIT press).
Niv, Y., Daw, N., Joel, D., and Dayan, P. (2006). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* 507–520.
Oades, R. (1985). The role of noradrenaline in tuning and dopamine in switching between signals in the cns. *Neurosci. Biobehav. Rev.* 9, 261–282.
Oudeyer, P. -Y., and Kaplan, F. (2006). Discovering communication. *Connect. Science* 18, 189–206.
Oudeyer, P. -Y., and Kaplan, F. (2007). What is intrinsic motivation? a typology of computational approaches. *Front. Neurorobot.* 1.
Oudeyer, P.-Y., Kaplan, F., and Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286.
Panksepp, J. (1998). Affective neuroscience: the Foundations of Human and Animal Emotions. (Oxford University Press).
Pettit, H., and Justice J. Jr. (1989). Dopamine in the nucleus accumbens during cocaine self-administration as studied by in vivo microdialysis. *Pharmacol. Biochem. Behav.* 34, 899–904.
Piaget, J. (1952). *The Origins of Intelligence in Children* (New York, NY, Norton).
Ploghaus, A., Tracey, I., Clare, S., Gati, J., Rawlins, J., and Matthews, P. (2000). Learning about pain: the neural substrate of the prediction error of aversive events. *Proc. Natl. Acad. Sci.* 97, 9281–9286.
Quaade, F., Vaernet, K., and Larsson, S. (1974). Stereotaxic stimulation and electrocoagulation of the lateral hypothalamus in obese humans. *Acta. Neurochir.* 30, 111–117.

91). Curious model-building control systems. In *Proceeding Inter-Conference on Neural Networks*, Singapore. IEEE, Vol. 2, PP.

redictive reward signal of dopamine neurons. *J. Neurophysiol.* 80,

ehavioral theories and the neurophysiology of reward. *Annu. Rev.* 115.

P., and Montague, P. (1997). A neural substrate of prediction and 275, 1593–1599.

3). A basic circuit for cortical organization. In Perspectives in Memory zzaniga ed. (MIT Press) PP. 93–134.

ecker, S., and Kapur, S. (2006). Dopamine, prediction error and ning: a model-based account. *Netw. Comput. Neural Sys.* 17,

e autotelic principle. In Embodied Artificial Intelligence, I. Fumiya, els and K. Kunyoshi eds. Vol. 3139 of *Lecture Notes in AI*, (Berlin, PP. 231–242.

e Neurobiology of Motivation and Reward (New York, NY, Springer

z, W. (2001). Temporal difference model reproduces anticipatory *Neural Comput.* 13, 841–862.

earning to predict by the methods of temporal differences. *Mach.*

, A. (1998). *Reinforcement Learning: An Introduction* (Cambridge,

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

Thelen, E., and Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action* (Boston, MA, USA, MIT Press).

Thorndike, E. (1911). *Animal Intelligence: Experimental Studies* (MacMillan, New York, NY).

Thrun, S., and Pratt, L. (1998). *Learning to Learn*. Kluwer Academic Publishers.

Walkins, C., and Dayan, P. (1992). Q-learning. *Mach. Learn.* 8, 279–292.

Weinberger, D. R. (1987). Implications of normal brain development for the pathogenesis of schizophrenia. *Arch. Gen. Psychiatry* 44, 660–669.

Weiner, I., and Joel, D. (2002). Dopamine in schizophrenia: dysfunctional information processing in basal ganglia-thalamocortical split circuits. In *Handbook of Experimental Pharmacology, vol 154/II, Dopamine in the CNS II*, G. Chiara ed. (Springer)PP. 417–472.

White, N. M. (1989). Reward or reinforcement: what's the difference? *Neurosci.Biobehav. Rev.* 13, 181–186.

White, R. (1959). Motivation reconsidered: The concept of competence. *Psychol. Rev.* 66, 297–333.

Wise, R. (1989). The brain and reward. In *The Neuropharmacological Basis of Reward*, J. Liebman and S. Cooper, eds. (Clarendon Press) PP. 377–424.

Yoshimoto, K., McBride, W., Lumeng, L., and Li, T.-K. (1991). Alcohol stimulates the release of dopamine and serotonin in the nucleus accumbens. *Alcohol.* 9, 17–22.

doi: 10.3389/neuro.01/1.1.017.2007