Computational Theories of Curiosity-Driven Learning

Pierre-Yves Oudeyer

▶ **To cite this version:**

Pierre-Yves Oudeyer. Computational Theories of Curiosity-Driven Learning. The New Science of Curiosity, 2018. hal-03513491

HAL Id: hal-03513491
https://inria.hal.science/hal-03513491v1

Submitted on 5 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:1802.10546v2 [cs.AI] 18 Jun 2018

Computational Theories of Curiosity-Driven Learning

Pierre-Yves Oudeyer*
Inria and Ensta ParisTech, France

Abstract

What are the functions of curiosity? What are the mechanisms of curiosity-driven learning? We approach these questions about the living using concepts and tools from machine learning and developmental robotics. We argue that curiosity-driven learning enables organisms to make discoveries to solve complex problems with rare or deceptive rewards. By fostering exploration and discovery of a diversity of behavioural skills, and ignoring these rewards, curiosity can be efficient to bootstrap learning when there is no information, or deceptive information, about local improvement towards these problems. We also explain the key role of curiosity for efficient learning of world models. We review both normative and heuristic computational frameworks used to understand the mechanisms of curiosity in humans, conceptualizing the child as a sense-making organism. These frameworks enable us to discuss the bi-directional causal links between curiosity and learning, and to provide new hypotheses about the fundamental role of curiosity in self-organizing developmental structures through curriculum learning. We present various developmental robotics experiments that study these mechanisms in action, both supporting these hypotheses to understand better curiosity in humans and opening new research avenues in machine learning and artificial intelligence. Finally, we discuss challenges for the design of experimental paradigms for studying curiosity in psychology and cognitive neuroscience.

Keywords: Curiosity, intrinsic motivation, world models, rewards, free-energy principle, learning progress hypothesis, lifelong learning, predictions, machine learning, AI, developmental robotics, development, curriculum learning, self-organization.

1. Introduction

Humans and many other animals spontaneously explore their environments. This often happens without a pressure for finding extrinsic rewards like food, and without external incentives from their social peers. Such spontaneous exploration seems to be produced by an internal mechanism pushing them to make sense of their world: they explore for the intrinsic purpose of getting better at predicting and controlling the world. This spontaneous investigation of the environment, and of the link between one's own physical and cognitive capabilities and the environment, can take many different forms. This ranges from babies trying various ways to locomote, or exploring grasping,

*http://www.pyoudeyer.com

manipulating, banging or throwing of all sorts of objects, or testing how social peers respond to vocalizations, to children practicing tool building with wooden sticks, or throwing wooden sticks in rivers to see how they will flow, to adults searching information about a hobby, learning a new sport, or a scientist turning his telescope towards faraway galaxies.

All these exploratory behaviours can be seen as questions posed by the organism about its environment or about the relation between its environment and it's own current state of knowledge and skills. These questions can be formulated in various ways ranging from actual physical/behavioural experimentation to formulating a linguistic question.

These mechanisms have been discussed from various perspectives in the scientific literature, and in particular using the related concepts of curiosity [Berlyne, 1960], intrinsic motivation [Harlow, 1950], and free play [Bruner et al., 1976]. Across many different fields, theorists have suggested that interest is engaged by what is just beyond current knowledge, neither too well known nor too far beyond what is understandable. This idea has been offered many times in psychology, through concepts like cognitive dissonance [Kagan, 1972], optimal incongruity [Hunt, 1965], intermediate novelty [Berlyne, 1960, Kidd et al., 2012] and optimal challenge [Csikszenthmihalyi, 1991], and recent research in neuroscience is now investigating the associated neural mechanisms [Gottlieb et al., 2013].

Several formal frameworks have recently enabled to improve our theoretical understanding of these mechanisms. This includes frameworks considering the curiosity system as a machine which goal is to build predictive and control models of the world [Kaplan and Oudeyer, 2007a, Oudeyer and Smith, 2016] - and sometimes the brain as a whole is conceptualized like this as in the free energy principle [Friston et al., 2017a, Friston et al., 2017b]. Related to this, reinforcement learning and optimization frameworks consider curiosity as a mechanism which allows to generate diversity in exploration, enabling to get out of local minima in searching for behaviours that maximize extrinsic rewards or fitness [Schmidhuber, 1991, Barto, 2013, Baldassarre and M., 2013, Lehman and Stanley, 2008, Bellemare et al., 2016, Forestier et al., 2017, Colas et al., 2018].

2. Curiosity for Exploration and Discovery in an Open World

An apparent evolutionary mystery is that such spontaneous investigations of the environment are often very costly in time and energy, and do not appear at first sight to provide a direct benefit for feeding, mating, survival and reproduction. So how could such mechanisms evolve? What is their function?

In an uncertain world with rare resources, one could expect that organisms spare their energy to explore only parts of the environment that are likely to provide information about where to get these resources. However, the real world is not only uncertain, but from the point of view of many basic physiological needs like finding food, it is full of multimodal multidimensional stimuli that are not obviously relevant to these needs. Animals have initially little ideas of what kinds of actions are required to fulfil them. Thus, when extrinsic rewards (resources) are hard to find, the main challenge is not to estimate the relative reward probabilities associated to a few reward-

relevant options in order to maximize the efficiency of a known solution. Rather, the challenge is to discover the first few bits of rewards and how to build coarse strategies to get them. In this context where discovery of new strategies and new outcomes becomes the main issue (as opposed to refining a known strategy for getting a known outcome), one can better see how to make sense of curiosity-driven exploration in living organisms.

Let's take the example of an 8-9 months old baby, sitting on the ground and alone in a room. He has seen a box of sweets on top of a kitchen's furniture, and aims to get them. While the baby might really be motivated to get the sweets for a moment, this task is for him a real conundrum. The situation is full of multiple kinds of deep uncertainties, and most importantly deep unknowns. First of all, the child has only a very approximate idea of the current state of the world: he sees the sweet box from far away with his eyes, and estimating simple things such as distance and height is already very difficult, since he is mostly used to interact with objects that are already in his hands, and has limited skills to estimate the state of far away objects. Second, the child has no clue about how to get to the sweet box, and has no clue where to look to find information about a solution. He does not even know how to stand up on its two feet, and his crawling strategy is very imprecise to move around. He does not know yet what a tool is, and his brain cannot even imagine at this point the possibility to push a chair next to the furniture, then try climb it to get to the box (at this point, chairs are perceived as obstacles for him). Here, he is much beyond uncertainty: it is meaningless for him to compute probabilities or uncertainties associated to the success of this strategy (and its associated sub-goals), as they involve events that are not already part of the space of hypotheses he can consider. Just think of the intermediate skill of standing up on its two feet and climbing the chair. Even if targeting these skills can be suggested by observing its social peers, they involve such complex internal muscular synergies that initially the child has also little cue about what patterns of muscle activations, and what proprioceptive and external visual information to attend to control these skills. Also very little can be inferred from observing others, as high-dimensional muscular activations and covert attention in these skills are not directly observable.

So what could the child do? He might consider a proxy internal measure for guiding its search for the sweet box, such as the current distance between his body and the sweets. Yet, optimizing it with its current skills would just make it crawl to the bottom of the furniture and extend its arm for a hopeless far reach. In that case this proxy measure would be less sparse than the sweet reward itself, but it would be highly deceptive, and equally inoperant. Rather, a good strategy for the child shall be to simply forget about this target sweet box for a while, go back to his playground and explore spontaneously, driven by curiosity, various forms of skills ranging from body locomotion to object manipulation and tool use. Playing around with its own body shall then lead him to discover both the concept of and strategies for climbing up furniture, as well as the concept and strategies for using objects of all kinds as tools. Then, when coming back a few months later to the task of getting to the sweet, and once he will have acquired a skill repertoire including walking, pushing chairs around, and using them as tools to get upwards, the solution might become much more accessible. At this point only the problem will be looking like a classical reinforcement learning problem in the lab with few obvious relevant options to choose from.

Research in developmental robotics and artificial intelligence has confirmed this intuition in

the last decade through computational and robotics experiments. Several strands of works have considered the problem of how a machine could solve difficult problems with rare or deceptive rewards [Barto et al., 2004, Schmidhuber, 1991, Lehman and Stanley, 2008], sometimes directly aiming at modelling how human children explore and learn in these contexts [Oudeyer et al., 2007, Oudeyer and Smith, 2016]. They found that various forms of curiosity-driven exploration can indeed be the key to make discoveries and solve these complex reinforcement learning problems.

An example is the artificial intelligence and robotics experiment presented in [Forestier et al., 2017] (see Figure 1), where a high-dimensional humanoid 'baby' robot is situated in an environment with many objects. Some of them are more or less difficult to learn to control, and some other are uncontrollable by the robot, such as another robot moving around. However, the 'baby' robot does not know initially which objects it can learn to control and which ones it cannot. Among these objects, a ball is placed in an area beyond the direct reach of the robot. Yet, other objects can be used as tools to enable the robot to move the ball. The robot can use its hand to push a joystick, which in turn moves a tele-operated crane toy, and this crane toy can push the ball. However, the robot has no concept of tool initially and does not know that these objects can physically interact: it has no pre-coded way to represent such physical interactions. It can send low-level motor commands into the motors of its arm. It can perceive object positions and movements individually. Yet, it does not know initially how specific sequences of motor commands relate to potential movements of each object (and how the movements of objects might relate to each other). While this robotic situation is simplified as compared to most real world situations encountered by human infants, it is already extremely complex. Indeed, the sequence of motor commands for a simple arm movement needed to reach for an object amounts to specifying several dozen continuous numbers, while the perception of the movement of each single object during one second is also encoded by several dozens continuous numbers. As a whole, the sensorimotor space that the robot explores has several hundred dimensions.

Given such an environment, let us imagine an engineer imposing the following external objective to the robot: it should learn to move the ball forward at a maximal speed. To define this objective, the engineer can design an extrinsic reward signal: each time the robot tries a sequence of motor commands that produce a movement of the ball, this reward is a scalar number proportional to the speed and target direction of the ball. Each time the motor commands do not produce any ball movement, the reward is zero.

The standard machine learning approach to enable the robot to solve this problem is reinforcement learning (RL) [Sutton and Barto, 2018]. This can be viewed as a family of optimization algorithms that learn optimal controllers, i.e. learn to produce sequences of motor commands that maximize the reward. The way standard approaches to RL work is through a combination of hill-climbing (gradient descent) and random exploration. For example, state-of-the-art deep reinforcement learning algorithms for learning continuous control, such as DDPG and related algorithms [Lillicrap et al., 2015, Schulman et al., 2017, Sigaud and Stulp, 2018], work by alternating between updating the current controller solution in order to climb the hill of rewards (this requires that rewards of different magnitudes are observed when slightly changing the controller), and producing random perturbations of the current best controller to obtain further information about the

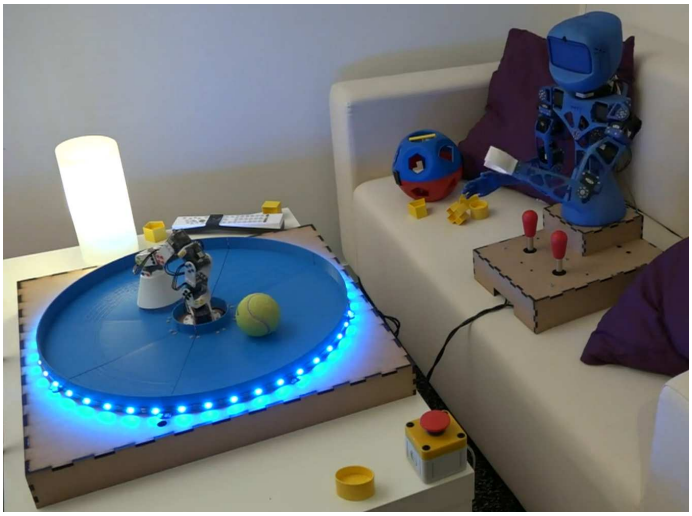

Figure 1. **Curiosity-driven exploration through autonomous goal setting and self-organized curriculum learning** in the experimental setup presented in [Forestier et al., 2017] (see video: `https://www.youtube.com/watch?v=NOLAwD4ZTW0`). Here a humanoid robot is surrounded by objects. The robot can learn to control some of them, while others are unlearnable distractors. Some objects can interact with each other, e.g. be used as tool to control other objects. The robot initially does not know which skills and objects are learnable, nor that objects may interact with each others. If an engineer would like the robot to learn how to move the ball forward by providing rewards only when the ball moves, then using classical reinforcement learning approaches would fail. Indeed, RL approaches combine hill-climing mechanisms that require observations of non-zero reward to improve the current solution, and rely on random exploration otherwise. Here, the time required by random exploration to enable the robot to find such a rare reward would be prohibitively long. The robot initially has no idea where to look for cues to solve this task. Indeed, moving the ball entails moving the white electric crane toy, which can only be moved by pushing one of the two joysticks, which can itself only be moved by appropriate movements of the hand, requiring specific sequences of motor commands in the joints. A more efficient approach is to use curiosity-driven exploration, where one lets the robot self-generate its own goals for controlling various objects, and spend time on the ones which produce maximal learning progress. The internal generation of goals, and the focus on goals providing maximal learning progress, model a form of curiosity. Doing so, the robot will first focus on playing with its hand, which is initially providing maximal learning progress. Then, it will discover as a side effect how to move the joysticks. In turn, due to the physical couplings in the environment, focusing on learning about the joysticks makes the robot discover how to move the crane toy and the ball in a few hours.

reward distribution.

However, such an RL approach will not work in the robotic environment considered here. Indeed, in this particular environment, the problem of moving the ball forward is said to be a 'rare reward' problem. Indeed, due to the structure of the environment, the vast majority of possible sequences of motor commands will produce a reward of zero. Actually, most random sequences of motor commands will make that the robot will not even touch the joystick in front of him. Thus, if the robot tries actions randomly, there is a very low probability to get a non-zero reward, i.e. that its arms touches the joystick in a peculiar way that makes the crane toy in a peculiar way that makes the ball move. This is a problem as the hill-climbing mechanisms of RL algorithms need to observe distributions of non-zero reward to update the controller, and doing random exploration of motor commands would here take a prohibitively long time before the first non-zero rewards are found, i.e. before finding the first few action sequences that produce ball movement. This problem of RL approaches that focus on hill-climing of the extrinsic reward is now well-known, and applies to man environments with rare or deceptive rewards[1] [Bellemare et al., 2016, Sigaud and Stulp, 2018, Colas et al., 2018].

An alternative computational approach studied in [Forestier et al., 2017] has consisted in equipping the robot with the capability to self-generate and explore many other goals than moving the ball. For example, the system can self-generate and focus on goals such as moving the hand in various directions, pushing any of the joysticks along particular trajectories, or trying to move the crane toy in diverse ways. The idea is that the system decides to spend time learning about one of these goals based on an intrinsic measure of the learning progress towards solving them, and ignoring completely how much they provide information about the goal of the engineer (move the ball). Formally, self-generating a goal (e.g. move toy in direction d) amounts to self-defining an internal reward function which quantifies how well sequence of motor commands produce the targeted outcome (e.g. the reward can be a scalar value proportional to the difference between the reached toy direction and the target toy direction). Then, the system can use RL algorithms to learn controllers for these internally defined goals and rewards. This whole process models several aspects of curiosity-driven exploration. First, it includes internal spontaneous generation of goals. Second, it includes a meta-cognitive mechanism to assess the relative 'interestingness' of self-generated goals, based on quantifying expected learning progress. This is used by the robot to decide on which goal to focus on at any given moment in time, and to self-organize a curriculum of learning.

Intuitively this may appear to be an even more difficult situation for the robot, as it consists in providing it with many other potentially difficult problems in addition to the initial problem of the engineer. However, this turned out to facilitate the acquisition of skills to move the ball forward. Indeed, due to the structure and richness of the physical environment, the robot manages to find

[1] Deceptive rewards are local improvement of rewards that push the learner to update the controller in wrong directions in relation with the global optimum. For example, in the infant example above, this would be the child observing that elevating the hand towards the sweet box on the table decreases the distance between him and the sweetbox. If the child is considering the hand-sweet distance as a measure of reward, this would be reinforcing this strategy. However, this would get the child away from more efficient strategies on the long term such as fetching a chair and climbing it.

continuously goals where learning can happen efficiently due to dense reward feedback, leading to an increase in behavioural diversity, and then to an increase of probability to find sequences of motor commands that solve other goals with rarer rewards. In this particular case, this curiosity-driven exploration mechanism pushed the robot to first focus on moving its hand around, leading to a greater diversity of motor behaviours. As a side effect, this enabled the discovery of how to move the joystick [2]. Then in turn, moving the joystick became interesting as a source of learning progress. This increased the focus on practicing goals related to the joysticks, which made the robot discover that the crane toy can be moved around, leading quickly to the discovery of how to move the ball. As the first ball movements are discovered, choosing the goal of moving the ball becomes easier, as the hill-climbing mechanism of reinforcement learning can efficiently compute how to change the controller to improve ball movements. Following this process, curiosity-driven exploration enables the robot to learn in a few hours how to solve the task of moving the ball around, even without initially pointing this task as particularly relevant among many other tasks that are also learnt. On the contrary, this would have been prohibitively long and difficult to learn by only considering and focusing on this task externally defined by the engineer.

This example relies on computational models of curiosity-driven exploration that were explicitly motivated by modeling mechanisms of human spontaneous exploration and their role in explaining the discovery of tool use [Forestier and Oudeyer, 2016b, Forestier and Oudeyer, 2016a], vocalization [Moulin-Frier et al., 2014] and language [Forestier et al., 2017] by children. However, several other strands of research in artificial intelligence have converged to the same conclusion that curiosity-driven exploration could be a key to solve complex tasks with rare or deceptive rewards. For example, several works in the field of evolutionary computation have shown that novelty search could be essential in solving difficult optimization problems [Lehman and Stanley, 2008], sometimes in combination with the task-specific reward [Mouret and Doncieux, 2012], but also sometimes by completely forgetting this task specific reward in the novelty search process [Lehman and Stanley, 2008]. In the domain of reinforcement learning, the idea of exploration bonuses [Andreae and Andreae, 1978, Sutton, 1990, Dayan and Sejnowski, 1996] has also been shown to increase the efficiency of reinforcement learning algorithms, with the addition of a reward component measuring quantities such as the novelty of a state [Sutton, 1990], prediction progress [Schmidhuber, 1991, Kaplan and Oudeyer, 2003], density of state exploration [Ostrovski et al., 2017], predictive information [Martius et al., 2013], predictive information gain [Little and Sommer, 2013], or empowerment [Salge et al., 2014]. Within an intrinsically motivated multi-goal reinforcement learning perspective, measures of competence progress towards self-generated goals were used to automatically generate learning curriculum for acquiring complex skills [Baranes and Oudeyer, 2013, Oudeyer et al., 2013, Nguyen and Oudeyer, 2014,

[2] When the robot targets a goal, but observes effects relevant to another goal as a side effect, it is able to learn retrospectively about this other goal using the collected observation. For example, if it targets to move the hand on the right, but in practice moves it on the left and moves the joystick, it uses this observation to learn how to move the ball left and how to move the joystick. This is a central property of these intrinsically motivated goal exploration processes [Baranes and Oudeyer, 2013, Forestier et al., 2017], which is shared with other related multi-goal learning algorithms in RL such as Hindsight Experience Replay [Andrychowicz et al., 2017].

Forestier et al., 2017]. Within a hierarchical reinforcement learning perspective, intrinsically motivated RL approaches have also been used as a framework enabling the discovery of hierarchically reusable skills for boosting exploration [Barto et al., 2004, Kulkarni et al., 2016, Machado et al., 2017]. Recent breakthrough in deep reinforcement learning have leveraged these various strands of works. For example, several deep reinforcement learning algorithms were shown to be able to solve difficult problems with rare rewards (e.g. video games), sometimes by even ignoring the score measure [Pathak et al., 2017], either by introducing an intrinsic reward measuring forms of novelty or learning progress [Ostrovski et al., 2017, Kulkarni et al., 2016], or by introducing auxiliary tasks [Jaderberg et al., 2016, Andrychowicz et al., 2017, Florensa et al., 2017, Cabi et al., 2017] in ways that are related to intrinsically motivated goal exploration approaches in developmental robotics [Baranes and Oudeyer, 2013, Forestier et al., 2017, Péré et al., 2018].

3. The Child as a Sense-Making Organism

This computational perspective provides a conceptual framework to understand how various forms of curiosity-driven exploration can be instrumental for an organism to discover solutions to extrinsic tasks that are essential to its survival, such as finding extrinsic rewards like food. In this context, it becomes possible to make sense of curiosity-driven exploration in an evolutionary perspective [Smith, 2013], and indeed further computational models have shown that intrinsic reward systems could be the result of evolutionary processes optimizing for an individual fitness to reproduce in changing environments [Singh et al., 2010]. This perspective converges with the hypothesis proposed in psychology that humans might be equipped with dedicated motivational neural circuitry pushing them to explore the world for the pure sake of making sense of it, and independently (yet complementarily) from the search of extrinsic rewards [Berlyne, 1960]. For example, Gopnick [Gopnik, 2012] has suggested that the child could be viewed as a curiosity-driven little scientist equipped with an intrinsic urge to discover the underlying causal structure of the world (see also [Chater and Loewenstein, 2016]).

Several theoretical models have considered mechanisms of curiosity-driven exploration as organized exploration of the world with the purpose to build good predictive and control models of the world dynamics [Oudeyer et al., 2007, Friston et al., 2017a]. For example, Friston and colleagues [Friston et al., 2017a] have developed a normative theoretical framework, termed the free energy principle, to characterize the theoretical properties of an ideal optimal Bayesian learner trying to build a good predictive model of the world. More precisely, the free energy principle framework views the brain as an active Bayesian inference system that aims to minimize future expected surprise. A useful property of this framework is to mathematically formalize various forms of uncertainties which are necessarily encountered by the active inference system, and which reduction can lead to corresponding various forms of exploration akin to curiosity (which are also present in the learning progress framework detailed below). For example, sources of uncertainty that appear when mathematically decomposing the expected free energy include uncertainty about hidden states given an action policy: minimizing it leads to active sensing and perceptual infer-

ence, i.e. a form of curiosity aiming to improve the subjective estimation of the current world state. Another dimension is uncertainty about policies in terms of expected future states or model parameters: this leads to forms of epistemic exploration and learning, i.e. a form of curiosity aiming to improve the predictive model of what will be observed in the future depending on one's own actions. Yet another dimension is uncertainty about the model structure itself: this leads to structure learning and insight, a form of curiosity aiming to find new abstractions with structures that enable better predictions of what is happening in the environment. While the concept of goals is not directly covered by the free-energy principle, other frameworks like the autonomous goal setting paradigm [Oudeyer and Kaplan, 2007, Forestier et al., 2017], presented in the robotic experiments above, point to other forms of uncertainty and curiosity centered around goals. Indeed, another form of uncertainty is related to the self-evaluation of goal competences: minimizing these forms of uncertainty lead to curiosity-driven self-generation and practice of goals to learn about one's own competence to achieve them.

4. Normative or Heuristic Accounts? The Learning Progress Hypothesis

As the landscape of computational/mathematical theories of curiosity-driven exploration quickly develops [Baldassarre and M., 2013, Oudeyer and Smith, 2016], one becomes better equipped with formal and experimental tools to conceptualize better what curiosity might be, and what it could be used for in humans. However, this landscape of theories also raises multiple open questions to explain human curiosity. A first fundamental question is how to articulate normative approaches (e.g. the free-energy principle, empowerment) with heuristics-based approaches (e.g. the learning progress hypothesis [Kaplan and Oudeyer, 2007a, Oudeyer and Smith, 2016]) to account for human behaviour and brain mechanisms. Taking the perspective of David Marr's levels of analysis for understanding cognition and the brain, shall normative approaches be standing at the computational level (describing which are the problems and the quantities that the brain actually target) and heuristic approaches at the algorithmic level (describing which algorithms can actually solve these problems) on top of the implementation level investigating how neural circuitry could implement these algorithms? The answer to this epistemological question does not appear to be resolved yet.

Indeed, while normative approaches like the free-energy principle have been used successfully to account for several behavioural and neural observations ranging from saccadic eye movements to habit learning to place cell activity [Friston et al., 2017a, Friston et al., 2017b], it relies on assumptions that are still speculative about what could be happening in the brain. First, it relies on the assumption that the human brain is capable to achieve (approximate) hierarchical Bayesian inference, but several strands of work have shown a number of situations where the brain uses heuristics that are far away from a Bayesian behaviour (e.g; [Gigerenzer and Goldstein, 1996]). Second, such a normative Bayesian framework requires that modelled human subjects initially know the full space of possible hypotheses about possible world states, possible policies, possible model parameters, and possible model structures. This deviates strongly from the deep unknown encountered

by children as described in the example above: new hypotheses can be formed out of interaction with the world and unsupervised representation learning. Finally, active structured Bayesian inference is computationally very costly [Buckley et al., 2017], and quickly become computationally intractable for problems of moderate size.

The challenges to address efficient and tractable curiosity-driven exploration in real world situations also underly the development of heuristic theories of curiosity-driven learning [Oudeyer et al., 2013]. Within the perspective of the 'child as a sense making organism', these heuristic theories have considered what mechanisms could enable efficient exploration and learning in high-dimensional physical bodies, and under limited resources of time, energy and cognitive capabilities. One such heuristic-oriented theoretical framework is the 'learning progress (LP) hypothesis', proposed in [Oudeyer and Kaplan, 2006, Kaplan and Oudeyer, 2007a, Oudeyer et al., 2007, Oudeyer and Kaplan, 2007]. Here, curiosity-driven exploration in living organisms is viewed as driven by the **heuristic** estimation and search of various forms of expected learning progress[3]. More precisely, this includes a heuristic meta-cognitive mechanism that estimates expected learning progress associated to competing activities (stimuli to observe, situations or self-generated goals to engage in). This estimation of LP is then used to choose which activities to explore, selecting in priority activities that maximize expected learning progress.

The fundamental similarity between the free-energy principle and the LP hypothesis is that both frameworks consider curiosity-driven exploration as a process that aims to collect new information to maximize the quality of an internal world model, and where this world model includes a model of self-knowledge and self-competences. Another similarity is that both frameworks consider various forms of learning progress, as the organism can learn various forms of knowledge and skills at various scales of space and time. Some forms of learning progress can result from an attentional action on a short time scale, providing information gain to better estimate the current world state. But it can be also result from the choice to focus on an activity for a longer time scale, producing various forms of improvement of a predictive world model, ranging from reduction in empirical prediction errors to reduction of uncertainty about these predictions (uncertainty could improve without improving the average prediction error, and still this would be a form of learning progress). This latter form of learning progress (longer time scale of learning, improvement of predictive world model) has been the focused of most computational experiment of the LP hypothesis so far [Oudeyer and Smith, 2016].

There are also fundamental differences between the free-energy principle and the LP framework. In the free-energy framework, the mechanisms used to represent and update world models and their associated uncertainties are based on Bayesian inference. On the contrary, the LP frame-

[3] Various works in artificial intelligence, machine learning and optimal experiment design have studied, in the last 50 years, how mechanisms of active learning can push a *machine* to explore parts of the state-action space that maximize forms of information gain and learning progress. However, these lines of work did not propose and study the hypothesis that related mechanisms could explain aspects of *human* curiosity-driven learning, and in other animals. Coming from this modeling perspective, the LP hypothesis was developed independantly of these lines of work in machine learning and AI. The convergence of these various strands of work towards related algorithms supports the strength and scope of these mechanisms.

work has considered heuristic algorithms to learn world models from observations, and to estimate uncertainty and learning progress, using mechanisms such as memory-based and lazy learning techniques [Forestier et al., 2017], as well as neural networks [Kaplan and Oudeyer, 2003] and evolution strategies [Benureau and Oudeyer, 2016]. One particular aspect of this difference is that heuristic-based approaches do not require that the learner knows all possible events, event representations, and model representations as it discovers the world (on the contrary, the normative framework requires priors about these events and their representations, which is untractable for real world situations). Unsupervised neural network representation learning techniques can for example be used to learn new spaces of representations in which to make predictions and set goals, as the world is being discovered (e.g. [Péré et al., 2018]). Unsupervised learning techniques are also used in the LP framework to learn incrementally abstractions of the low-level sensorimotor flow, enabling for example to form distinct concepts of the self, of inanimate objects and of others based on their associated learnability properties [Oudeyer et al., 2007]. Finally, another difference already mentionned earlier, is that the LP framework has been extended to cover mechanisms of autonomous goal setting, which is a key dimension of curiosity-driven exploration in humans.

As the LP framework has studied what heuristics can drive efficient learning of world models in the real world, evaluation has leveraged robotics experiments under constraints of time and processing, showing how these mechanisms enable learning multiple forms of locomotion [Baranes and Oudeyer, 2013], manipulation of flexible objects [Nguyen and Oudeyer, 2014], and tool use discovery [Forestier et al., 2017] in high-dimensional continuous action and perceptual spaces.

Beyond its theoretical origin, the learning progress hypothesis makes behavioural predictions that are compatible with a growing set of experimental evidences in psychology. For example, Gerken et al. [Gerken et al., 2011] showed that 17-months old children attend more to learnable artificial linguistic patterns than to unlearnable ones. Also, Kidd et al. [Kidd et al., 2012] showed that infants prefer sequences of perceptual stimuli that have an intermediate level of complexity. Similarly, Begus et al. [Begus et al., 2016] showed that infants selectively ask the help of informants based on expected information they can provide. Baranes et al. [Baranes et al., 2014] showed how adult subjects, who were free to explore and select tasks of various difficulties, spontaneously organize their exploration in growing order of difficulty and settle on levels of intermediate difficulty just beyond their current skill level.

5. Curiosity and Self-Organization in Development

These computational studies of the learning progress hypothesis have also uncovered a crucial emergent property of such a heuristic mechanism: it spontaneously leads the learner to avoid situations which are either trivial or too complicated[4], and focus on situations that are just beyond

[4]Other heuristics like novelty or surprise bonuses proposed in the RL litterature do not scale as well in open real world environments as there are many tasks or situations which produce novelty or surprise but yet should be avoided as they are not learnable. An activity can be unlearnable due to unobservable causal factors, or to intrinsic unpredictability.

the current skills in prediction or control, exploring them as long as learning progress happens in practice.

This has enabled to generate the new hypothesis that mechanisms of curiosity drive the emergence of ordered developmental trajectories at long time scales [Kaplan and Oudeyer, 2007a, Oudeyer and Smith, 2016]. Several studies have shown that such trajectories match several fundamental properties of infant trajectories in domains such as vocal development [Moulin-Frier et al., 2014], imitation [Kaplan and Oudeyer, 2007b] and tool use discovery [Forestier and Oudeyer, 2016a, Forestier and Oudeyer, 2016c]. Related models of curiosity-driven learning, with intrinsic rewards based on information gain measured through empirical prediction errors, have also shown how it could model the formation of pro-social behaviors [Baraglia et al., 2016], or model the development of aspects of binocular vision [Zhu et al., 2017] or visual search in infants [Twomey and Westermann, 2017], as well as different forms of exploratory behaviours in other animals [Gordon et al., 2014].

5.1. The Playground Experiment

An example of self-organization of structured developmental trajectories driven by curiosity-driven exploration is the Playground Experiment [5] [Oudeyer and Kaplan, 2006, Kaplan and Oudeyer, 2007a, Oudeyer et al., 2007]. It illustrates how mechanisms of curiosity-driven exploration, dynamically interacting with learning, physical and social constraints, can self-organize developmental trajectories. In particular, this leads a learner to successively discover two important functionalities: object affordances and vocal interaction with its peers.

In these Playground Experiments, a quadruped learning robot (the learner) is placed on an infant play mat with a set of nearby objects and is joined by an adult robot (the teacher), see Figure 2 (A) [Oudeyer and Kaplan, 2006, Kaplan and Oudeyer, 2007a, Oudeyer et al., 2007]. On the mat and near the learner are objects for discovery: an elephant (which can be bitten or grasped by the mouth), a hanging toy (which can be bashed or pushed with the leg). The teacher is preprogrammed to imitate (with a different pitch of voice) the sounds made by the learner when the learning robot looks to the teacher while vocalizing at the same time.

The learner is equipped with a repertoire of motor primitives parameterized by several continuous numbers that control movements of its legs, head and a simulated vocal tract. Each motor primitive is a dynamical system controlling various forms of actions: (a) turning the head in different directions; (b) opening and closing the mouth while crouching with varying strength and timing; (c) rocking the leg with varying angle and speed; (d) vocalizing with varying pitch and length. These primitives can be combined to form a large continuous space of possible actions. Similarly, sensory primitives allow the robot to detect visual movements, salient visual properties, proprioceptive touch in the mouth, and pitch and length of perceived sounds. For the robot, these motor and sensory primitives are initially black boxes and he has no knowledge about their semantics, effects or relations.

[5] The text describing the Playground Experiment in this section, and the interaction of curiosity with social guidance below, is partly adapated from [Oudeyer and Smith, 2016].

The learning robot learns how to use and tune these primitives to produce diverse effects on its surrounding environment, and exploration is driven by the maximization of learning progress: the robot chooses to perform physical experiences (experiments) that improve maximally the quality of predictions of the consequences of its actions, i.e. that improve its world model.

Figure 2 (B) outlines the computational architecture of curiosity-driven learning, called IAC, used in the playground experiment [Oudeyer et al., 2007]. A prediction machine (M) learns to predict the consequences of actions taken by the robot in given sensory contexts. For example, this module might learn to predict (with a neural network) which visual movements or proprioceptive perceptions result from using a leg motor primitive with certain parameters. A metacognitive module estimates the evolution of errors in prediction of M in various regions of the sensorimotor space. More precisely, this module estimates the decrease in prediction errors for particular kinds of actions or particular contexts. An example of such a context could be the prediction of the consequences of a leg movement when this action is applied towards a particular area of the environment. These estimates of error reduction, measuring a form of learning progress, are used to compute an intrinsic reward. This reward is an internal quantity (a number) that is proportional to the decrease of prediction errors, and the maximization of this quantity is the objective of action selection within a computational reinforcement learning architecture [Kaplan and Oudeyer, 2003, Oudeyer et al., 2007, Oudeyer and Kaplan, 2007]. Importantly, the action selection system chooses most often to explore activities where the expected intrinsic reward is high, i.e. where the expected learning progress is high. However, this choice is probabilistic, which leaves the system open to learning in new areas and open to discovering other activities that may also yield progress in learning[6]. Since the sensorimotor flow does not come pre-segmented into activities and tasks, a system that seeks to maximize differences in learnability is also used to progressively categorize the sensorimotor space into differentiated contexts. This categorization thereby models the incremental creation and refining of cognitive categories differentiating activities/tasks.

To illustrate how such an exploration mechanism can automatically generate ordered learning stages, let us first imagine a learner confronted with four categories of activities, as shown on figure 2 (C). The practice of each of these four activities, which can be of varying difficulty, leads to different learning rates at different points in time (see the top curves, which show the evolution of prediction errors in each activity if the learner were to focus full-time and exclusively on each). If, however, the learner uses curiosity-driven exploration to decide what and when to practice by focusing on progress niches, it will avoid activities already predictable (curve 4) or too difficult to learn to predict (curve 1), in order to focus first on the activity with the highest learning rate (curve 3) and eventually, when the latter starts to reach a 'plateau', to switch to the second most promising

[6]Technically the decision on how much time to spend on a given activity/context is achieved using Multi-Armed Bandit algorithms for the so-called exploration/exploitation dilemma [Audibert et al., 2009]. As the measure of learning progress in each arm is used as the reward to maximize, this is a non-stationary bandit algorithm setting. However, a specificity of learning architectures used in the robotic experiments presented here is that instead of relying on a set of pre-defined bandit arms [Audibert et al., 2009], an unsupervised learning algorithm dynamically builds new bandit arms to select from [Baranès and Oudeyer, 2009].

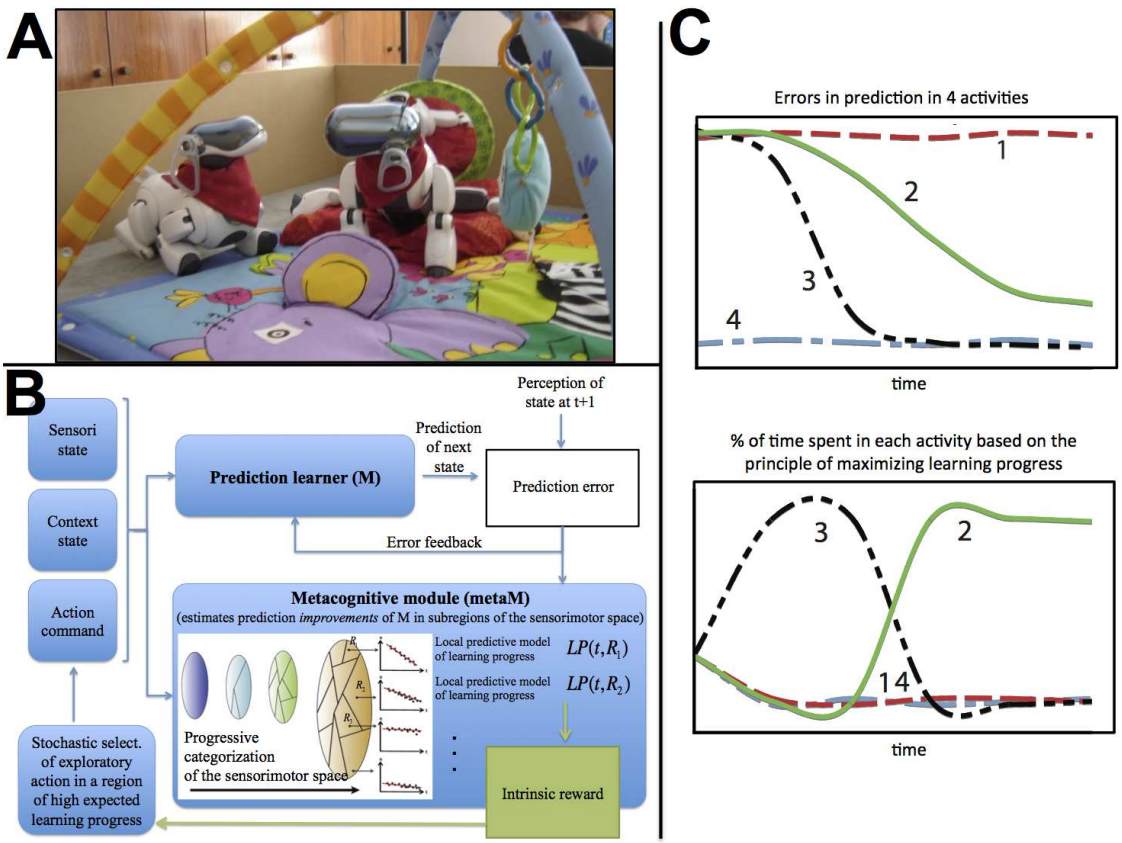

Figure 2. The Playground Experiment [Oudeyer and Kaplan, 2006, Oudeyer et al., 2007] (A) The learning context; (B) The computational architecture for curiosity-driven exploration: 1) the robot learner probabilistically selects actions and contexts according to their potential to provide information that improves the world model (i.e. reduces prediction errors); 2) an unsupervised learning algorithms progressively differentiates actions and contexts to be selected; (C) Illustration of a self-organized developmental sequence where the robot automatically identifies, categorizes and shifts from simple to mode complex learning experiences. Figure adapted with permission from [Gottlieb et al., 2013].

learning situation (curve 2). Thus, embodied exploration driven by learning progress creates an organized exploratory strategy, i.e. a developmental trajecotory: the system systematically achieves these learning experiences in an order and does so because they yield (given the propensities of the learner and the physical world) different patterns of uncertainty reduction.

In the Playground experiment, multiple experimental runs lead to two general categories of results: self-organization and a mixture of regularities and diversities in the developmental patterns

[Oudeyer and Kaplan, 2006, Oudeyer et al., 2007].

5.2. Self-Organization

In all of the runs, one observes the self-organization of structured developmental trajectories, where the robot explores objects and actions in a progressively more complex stage-like manner. During this exploration, the robot acquires autonomously diverse affordances and skills that can be reused later on and that change the learning progress in more complicated tasks, triggering a developmental cascade. The following developmental sequence was typically observed:

1. In a first phase, the learner achieves unorganized body babbling;
2. In a second phase, after learning a first rough model and meta-model[7], the robot stops combining motor primitives, exploring them one by one, but each primitive is explored itself in a random manner;
3. In a third phase, the learner begins to experiment with actions towards zones of its environment where the external observer knows there are objects (the robot is not initially provided with a representation of the concept of object), but in a non-affordant manner (e.g. it vocalizes at the non-responding elephant or tries to bash the teacher robot which is too far to be touched);
4. In a fourth phase, the learner now explores the affordances of different objects in the environment: typically focussing first on grasping movements with the elephant, then shifting to bashing movements with the hanging toy, and finally shifting to explorations of vocalizing towards the imitating teacher.
5. In the end, the learner has learnt sensorimotor affordances with several objects, as well as social affordances, and has mastered multiple skills. None of these specific objectives were pre-programmed. Instead, they self-organized through the dynamic interaction between curiosity-driven exploration, statistical inference, the properties of the body, and the properties of the environment.

These playground experiments do not simply simulate particular skills (such as batting at toys to make them swing or vocalizations) but simulate an ordered and systematic developmental trajectory, with a universality and stage-like structure that may be mistakenly taken to indicate an internally-driven process of maturation. However, the trajectory is created through activity and through the general principle that sensorimotor experiences that reduce uncertainty in prediction are rewarding. In this way, developmental achievements can build on themselves without specific pre-programmed dependencies but nonetheless like evolution itself create structure (see [Smith and Breazeal, 2007, Smith, 2013], for related findings and arguments).

[7]The 'model' refers here to the predictive world model being learnt, which enables to predict the consequences of actions in a given context. The 'meta-model' is another model built by the meta-cognite process, and continuously estimates expected learning progress of the lower-level world model.

5.3. Regularities and Diversity

Because these are self-organizing developmental processes, they generate not only strong regularities but also diversity across individual developmental trajectories. For example, in most runs one observes successively unorganized body babbling, then focused exploration of head movements, then exploration of touching an object, then grasping an object, and finally vocalizing towards a peer robot (pre-programmed to imitate). This can be explained as a gradual exploration of new progress niches, and those stages and their ordering can be viewed as a form of attractor in the space of developmental trajectories. Yet, with the same mechanism and same initial parameters, individual trajectories may invert stages, or even generate qualitatively different behaviours. This is due to stochasticity (the same motor commands do not produce always the same results), to small variability in the physical realities and to the fact that this developmental dynamical system has several attractors with more or less extended and strong domains of attraction (an attractor is a part of the state-space in which the dynamical system converges, depending on what was his initial state). We see this diversity as a positive outcome since individual development is not identical across different individuals but is always, for each individual, unique in its own ways. This kind of approach, then, offers a way to understand individual differences as emergent in developmental processes itself and makes clear how developmental processes might vary across contexts, even with an identical learning mechanism.

A further result to be highlighted is the early development of vocal interaction. With a single generic mechanism, the robot both explores and learns how to manipulate objects and how to vocalize to trigger specific responses from a more mature partner [Oudeyer and Kaplan, 2006, Kaplan and Oudeyer, 2007a]. Vocal babbling and language games have been shown to be key in infant language development; however, the motivation to engage in vocal play has often been associated with hypothesized language specific motivation. The Playground Experiment makes it possible to see how the exploration and learning of communicative behaviour might be at least partially explained by general curiosity-driven exploration of the body affordances, as also suggested by Oller [Oller, 2000]. Exploring this idea further, Forestier and Oudeyer [Forestier and Oudeyer, 2017] studied simulation showing how these mechanisms can drive the joint development of speech and tool use, where speech is discovered as a particular tool enabling to get social peers achieve targeted actions.

5.4. Interaction with Social Guidance

Other robotic models have explored how social guidance can be leveraged by an intrinsically motivated active learner and dynamically interact with curiosity to structure developmental trajectories [Thomaz and Breazeal, 2008, Nguyen and Oudeyer, 2014]. Focusing on vocal development, Moulin-Frier et al. conducted experiments where a robot explored the control of a realistic model of the vocal tract in interaction with vocal peers through a drive to maximize learning progress [Moulin-Frier et al., 2014]. This model relied on a physical model of the vocal tract, its motor control and the auditory system. The experiments showed self-organization of vocal development

trajectories that share structural similarities with infants [Oller, 2000]. They showed how these mechanism generate an adaptive transition from vocal self-exploration with little influence from the speech environment, to a later stage where vocal exploration becomes influenced by vocalizations of peers. Within the initial self-exploration phase, a sequence of vocal production stages self-organizes, and shares properties with infant data: the vocal learner first discovers how to control phonation, then vocal variations of unarticulated sounds, and finally articulated proto-syllables. As the vocal learner becomes more proficient at producing complex sounds, the vocalizations of the teacher become vocal goals to imitate that provide high learning progress, resulting in a shift from self-exploration to vocal imitation.

6. Challenges and Perspectives

Computational theories have enabled to better understand the potential structures and functions of curiosity-driven learning in the last decade. These theories have identified a wide diversity of algorithmic mechanisms that could produce the kind of spontaneous exploration displayed by humans and other animals. This diversity concerns both the measures of interests (e.g. novelty, surprise, learning progress, knowledge gap, intermediate complexity, ...) and the entities to which the brain may apply them (e.g. actions, states, goals, objects, tools, places, games, activities, learning strategies, social informants, ...), with time scales ranging from the moment-to-moment to days and months. Furthermore, theoretical models of curiosity-driven learning, and their application in artificial intelligence and machine learning, have shown the key role of these mechanisms for making discoveries and solving real-world problems with rare or deceptive rewards, in large and changing environments. In brief, computational theories:

1. have shown that the term 'curiosity' covers a wide diversity of complex mechanisms, generating different forms of exploration;

2. have proposed ways to model and study these mechanisms formally, contributing to the naturalization of the concept of 'curiosity';

3. have shown that curiosity mechanism are essential to learning and development, and thus should become a central topic in cognitive sciences and neurosciences.

Related work in psychology and neuroscience are beginning to converge with computational theories towards conceptualizing how mechanisms of curiosity can play a fundamental role in many aspects of development, ranging from sensorimotor, cognitive, emotional, to social development. However, for multiple reasons, experimental work studying the underlying mechanisms of curiosity have been very limited so far in psychology and neuroscience. The empirical testing of computational theories have been for a large part beyond the reach of existing experimental paradigms in psychology and neuroscience. Several challenges need to be addressed to leverage further the interaction between theory and experimentation.

The need for novel experimental paradigms in psychology and cognitive neuroscience. A first general challenge is that curiosity denotes a set of mechanisms that push individuals to explore what is interesting for themselves, out of the consideration of external tasks or expectations of social peers. Yet, the very act of participating to an experiment in a lab brings up expectations in the subject's mind about what the experimenter wants to observe or analyze, or will think about what they do. In the lab, curiosity can disappear quickly as soon as one begins to observe it. This is probably less the case with very young infants, but in their case the presence of social peers is also bound to influence what they do, and their limited capabilities for physical exploration and verbal reporting makes it difficult to study advanced forms of curiosity. So, how to study curiosity when setting up a controlled experiment introduces complex interaction with other motivational forces that are hard to control and evaluate? It is interesting to note that the most clear observations of curiosity in the lab do not come from studies targeting curiosity and information-seeking, but are rather observations of child behaviour spontaneously doing things that are wildly different from the task the experimenter designed for them. For example, in recent experiments of Lauriane Rat-Fisher and colleagues[8] about tool use development, children are asked to retrieve a salient toy stuck in a tube (the toy was expected to be very attractive to the child). Yet, several children showed spontaneous strong intrinsic interest in exploring how to push sticks and objects in the tube, continuing to do it with a lot of fun after getting the toy out of the tube, and completely ignoring the toy. Unfortunately, these making off observations are typically removed from the lens of analysis of traditional experimental studies, while they may display some of the most fundamental and mysterious mechanisms of learning and cognition.

Another experimental challenge is how to disentangle the potentially different mechanisms identified by theory, which may be simultaneously at play in individuals, and potentially on different time scales. For example, it could be possible that curiosity-driven attention on very short time scales may be driven by intrinsic rewards measuring different forms of novelty, surprise or prediction error. However, on longer time-scales, the curious brain may value the intrinsic interest of activities, games or goals with other measures of interest like learning progress. The study of curiosity over long time scales, focusing on how it may contribute to sculpt sensorimotor, cognitive and social development, on how it develops itself with time, and on how it interacts with other developmental forces such as social learning, is maybe the most important and most difficult challenge in this scientific area.

7. Acnknowledgements

This manuscript has benefited from very useful feeback and discussions with members of the Flowers team at Inria, as well as with Jacqueline Gottlieb, Linda Smith and Olivier Sigaud.

[8] personal communication

References

[Andreae and Andreae, 1978] Andreae, P. M. and Andreae, J. H. (1978). A teachable machine in the real world. *International Journal of Man-Machine Studies*, 10(3):301–312.

[Andrychowicz et al., 2017] Andrychowicz, M., Crow, D., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. (2017). Hindsight experience replay. In *Advances in Neural Information Processing Systems*.

[Audibert et al., 2009] Audibert, J. Y., Munos, R., and Szepesvri, C. (2009). Explorationexploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

[Baldassarre and M., 2013] Baldassarre, G. and M., M. (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer.

[Baraglia et al., 2016] Baraglia, J., Nagai, Y., and Asada, M. (2016). Emergence of altruistic behavior through the minimization of prediction error. *IEEE Transactions on Cognitive and Developmental Systems*, 8(3):141–151.

[Baranes et al., 2014] Baranes, A., Oudeyer, P., and Gottlieb, J. (2014). The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. *Frontiers in Neuroscience*.

[Baranès and Oudeyer, 2009] Baranès, A. and Oudeyer, P.-Y. (2009). R-iac: Robust intrinsically motivated exploration and active learning. *IEEE Transactions on Autonomous Mental Development*, 1(3):155–169.

[Baranes and Oudeyer, 2013] Baranes, A. and Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1).

[Barto, 2013] Barto, A. (2013). Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems*, pages 17–47. Springer Berlin Heidelberg.

[Barto et al., 2004] Barto, A. G., Singh, S., and Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd International Conference on Development and Learning*, pages 112–19.

[Begus et al., 2016] Begus, K., Gliga, T., and Southgate (2016). Infants' preferences for native speakers are associated with an expectation of information. *Proc Natl Acad Sci U S A*, 113:12397–12402.

[Bellemare et al., 2016] Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., , and Munos, R. (2016). Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, page 14711479.

[Benureau and Oudeyer, 2016] Benureau, F. C. and Oudeyer, P.-Y. (2016). Behavioral diversity generation in autonomous exploration through reuse of past experience. *Frontiers in Robotics and AI*, 3:8.

[Berlyne, 1960] Berlyne, D. (1960). *Conflict, Arousal and Curiosity*. McGraw-Hill, New York.

[Bruner et al., 1976] Bruner, J. S., Jolly, A., and Sylva, K. (1976). *Play: Its role in development and evolution*. Basic Books.

[Buckley et al., 2017] Buckley, C., Kim, C., McGregor, S., and Seth, A. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79.

[Cabi et al., 2017] Cabi, S., Colmenarejo, S. G., Hoffman, M. W., Denil, M., Wang, Z., and De Freitas, N. (2017). The intentional unintentional agent: Learning to solve many continuous control tasks simultaneously. In *1st Conference on Robot Learning*, CoRL.

[Chater and Loewenstein, 2016] Chater, N. and Loewenstein, G. (2016). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization*, 126:137–154.

[Colas et al., 2018] Colas, C., Sigaud, O., and Oudeyer, P.-Y. (2018). Gep-pg: Decoupling exploration and exploitation in deep reinforcement learning algorithms. In *International Conference on Machine Learning (ICML)*.

[Csikszenthmihalyi, 1991] Csikszenthmihalyi, M. (1991). *Flow-the Psychology of Optimal Experience*. Perennial, Harper.

[Dayan and Sejnowski, 1996] Dayan, P. and Sejnowski, T. J. (1996). Exploration bonuses and dual control. *Machine Learning*, 25(1):522.

[Florensa et al., 2017] Florensa, C., Held, D., Wulfmeier, M., Zhang, M., and Abbeel, P. (2017). Reverse curriculum generation for reinforcement learning. In *Proceedings of the 1st Annual Conference on Robot Learning, in PMLR 78*, pages 482–495.

[Forestier et al., 2017] Forestier, S., Mollard, Y., and Oudeyer, P.-Y. (2017). Intrinsically motivated goal exploration processes with automatic curriculum learning. *arXiv preprint arXiv:1708.02190*.

[Forestier and Oudeyer, 2016a] Forestier, S. and Oudeyer, P.-Y. (2016a). Curiosity-driven development of tool use precursors: a computational model. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pages 1859–1864.

[Forestier and Oudeyer, 2016b] Forestier, S. and Oudeyer, P.-Y. (2016b). Modular active curiosity-driven discovery of tool use. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 3965–3972. IEEE.

[Forestier and Oudeyer, 2016c] Forestier, S. and Oudeyer, P.-Y. (2016c). Overlapping waves in tool use development: a curiosity-driven computational model. In *The Sixth Joint IEEE International Conference Developmental Learning and Epigenetic Robotics*.

[Forestier and Oudeyer, 2017] Forestier, S. and Oudeyer, P.-Y. (2017). A unified model of speech and tool use early development. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.

[Friston et al., 2017a] Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017a). Active inference: A process theory. *Neural computation*, 29(1):149.

[Friston et al., 2017b] Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. (2017b). Active inference, curiosity and insight. *Neural computation*, 29(10):2633–2683.

[Gerken et al., 2011] Gerken, L., Balcomb, F. K., and Minton, J. L. (2011). Infants avoid labouring in vain by attending more to learnable than unlearnable linguistic patterns. *Developmental science*, 14(5):972–979.

[Gigerenzer and Goldstein, 1996] Gigerenzer, G. and Goldstein, G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650.

[Gopnik, 2012] Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, 337(6102):1623–1627.

[Gordon et al., 2014] Gordon, G., Fonio, E., and Ahissar, E. (2014). Emergent exploration via novelty management. *Journal of Neuroscience*, 34(38):12646–12661.

[Gottlieb et al., 2013] Gottlieb, J., Oudeyer, P.-Y., Lopes, M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593.

[Harlow, 1950] Harlow, H. (1950). Learning and satiation of response in intrinsically motivated complex puzzle performances by monkeys. *J. Comp. Physiol. Psychol*, 43:289–294.

[Hunt, 1965] Hunt, J. (1965). Intrinsic motivation and its role in psychological development. In *Nebraska Symposium on Motivation*, volume 13, pages 189–282.

[Jaderberg et al., 2016] Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. (2016). Reinforcement learning with unsupervised auxiliary tasks. arXiv preprint arXiv:1611.05397.

[Kagan, 1972] Kagan, J. (1972). Motives and development. *J. Pers. Soc. Psychol*, 22:5166.

[Kaplan and Oudeyer, 2003] Kaplan, F. and Oudeyer, P.-Y. (2003). Motivational principles for visual know-how development. In Prince, C., Berthouze, L., Kozima, H., Bullock, D., Stojanov, G., and Balkenius, C., editors, *Proceedings of the 3rd international workshop on Epigenetic Robotics : Modeling cognitive development in robotic systems, no. 101*, pages 73–80. Lund University Cognitive Studies.

[Kaplan and Oudeyer, 2007a] Kaplan, F. and Oudeyer, P.-Y. (2007a). In search of the neural circuits of intrinsic motivation. *Frontiers in neuroscience*, 1:17.

[Kaplan and Oudeyer, 2007b] Kaplan, F. and Oudeyer, P.-Y. (2007b). The progress-drive hypothesis: an interpretation of early imitation. In Dautenhahn, K. and Nehaniv, C., editors, *Models and mechanisms of imitation and social learning: Behavioural, social and communication dimensions*, pages 361–377. Cambridge University Press.

[Kidd et al., 2012] Kidd, C., Piantadosi, S., and Aslin, R. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLOS ONE*, 7:e36399.

[Kulkarni et al., 2016] Kulkarni, T., Narasimhan, K., Saeedi, A., , and Tenenbaum, J. (2016). Hierarchical deep reinforce- ment learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in Neural Information Processing Systems*, page 36753683.

[Lehman and Stanley, 2008] Lehman, J. and Stanley, K. O. (2008). Exploiting open-endedness to solve problems through the search for novelty. In *ALIFE*, pages 329–336.

[Lillicrap et al., 2015] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

[Little and Sommer, 2013] Little, D. Y. and Sommer, F. T. (2013). Learning and exploration in action-perception loops. *Frontiers in neural circuits, 7*.

[Machado et al., 2017] Machado, M. C., Bellemare, M. G., and Bowling, M. (2017). A laplacian framework for option discovery in reinforcement learning. In *International Conference on Machine Learning (ICML)*.

[Martius et al., 2013] Martius, G., Der, R., and Ay, N. (2013). Information driven self-organization of complex robotic behaviors. *PloS one*, 8(5):e63400.

[Moulin-Frier et al., 2014] Moulin-Frier, C., Nguyen, S. M., and Oudeyer, P. Y. (2014). Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Frontiers in psychology*, 4:1006.

[Mouret and Doncieux, 2012] Mouret, J. B. and Doncieux, S. (2012). Encouraging behavioral diversity in evolutionary robotics: An empirical study. *Evolutionary computation*, 20(1):91–133.

[Nguyen and Oudeyer, 2014] Nguyen, M. and Oudeyer, P.-Y. (2014). Socially guided intrinsic motivation for robot learning of motor skills. *Autonomous Robots*, 36(3):273–294.

[Oller, 2000] Oller, D. (2000). *The emergence of the speech capacity*. Lawrence Erlbaum and Associates Inc, Mahwah, NJ.

[Ostrovski et al., 2017] Ostrovski, G., Bellemare, M. G., Oord, A. V. D., and Munos, R. (2017). Count-based exploration with neural density models. In *Proceedings of the International Conference on Machine Learning*.

[Oudeyer and Kaplan, 2007] Oudeyer, P. and Kaplan, F. (2007). What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6.

[Oudeyer et al., 2013] Oudeyer, P.-Y., Baranes, A., and Kaplan, F. (2013). Intrinsically motivated learning of real-world sensorimotor skills with developmental constraints. In *Intrinsically motivated learning in natural and artificial systems*, page 303365. Springer, Berlin, Heidelberg.

[Oudeyer and Kaplan, 2006] Oudeyer, P.-Y. and Kaplan, F. (2006). Discovering communication. *Connection Science*, 18(2):189–206.

[Oudeyer et al., 2007] Oudeyer, P.-Y., Kaplan, F., and Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput*, 11(2):265–286.

[Oudeyer and Smith, 2016] Oudeyer, P.-Y. and Smith, L. (2016). How evolution can work through curiosity-driven developmental process. *Top. Cogn. Sci*, 8(2):492–502.

[Pathak et al., 2017] Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*.

[Péré et al., 2018] Péré, A., Forestier, S., Sigaud, O., and Oudeyer, P.-Y. (2018). Unsupervised learning of goal spaces for intrinsically motivated goal exploration. In *International Conference on Learning Representations (ICLR)*.

[Salge et al., 2014] Salge, C., Glackin, C., and Polani, D. (2014). Changing the environment based on empowerment as intrinsic motivation. *Entropy*, 16(5):2789–2819.

[Schmidhuber, 1991] Schmidhuber, J. (1991). Curious model-building control systems. In *Proceedings of the International Joint Conference on Neural Network*, volume 2, pages 1458–1463.

[Schulman et al., 2017] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[Sigaud and Stulp, 2018] Sigaud, O. and Stulp, F. (2018). Policy search in continuous action domains: an overview. *arXiv preprint arXiv:1803.04706*.

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

[Smith, 2013] Smith, L. (2013). Its all connected: Pathways in visual object recognition and early noun learning. *American Psychologist*, 68(8):618.

[Smith and Breazeal, 2007] Smith, L. and Breazeal, C. (2007). The dynamic lift of developmental process. *Developmental Science*, 10(1):61–68.

[Sutton, 1990] Sutton, R. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 216–224.

[Sutton and Barto, 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge. Second Edition.

[Thomaz and Breazeal, 2008] Thomaz, A. L. and Breazeal, C. (2008). Experiments in socially guided exploration: Lessons learned in building robots that learn with and without human teachers. *Connection Science*, 20(23):91–110.

[Twomey and Westermann, 2017] Twomey, K. and Westermann, G. (2017). Curiosity-based learning in infants: A neurocomputational approach. *Developmental Science*.

[Zhu et al., 2017] Zhu, Q., Triesch, J., and Shi, B. (2017). Joint learning of binocularly driven saccades and vergence by active efficient coding. *Frontiers in neurorobotics*, 11:58.