Review

Cell PRESS

Information-seeking, curiosity, and attention: computational and neural mechanisms

Jacqueline Gottlieb[1,2], Pierre-Yves Oudeyer[3,4], Manuel Lopes[3,4], and Adrien Baranes[1]

[1] Department of Neuroscience, Columbia University, New York, NY, USA
[2] Kavli Institute for Brain Science, Columbia University, New York, NY, USA
[3] Inria, Bordeaux, France
[4] Ensta ParisTech, Paris, France

Intelligent animals devote much time and energy to exploring and obtaining information, but the underlying mechanisms are poorly understood. We review recent developments on this topic that have emerged from the traditionally separate fields of machine learning, eye movements in natural behavior, and studies of curiosity in psychology and neuroscience. These studies show that exploration may be guided by a family of mechanisms that range from automatic biases toward novelty or surprise to systematic searches for learning progress and information gain in curiosity-driven behavior. In addition, eye movements reflect visual information searching in multiple conditions and are amenable for cellular-level investigations. This suggests that the oculomotor system is an excellent model system for understanding information-sampling mechanisms.

Information-seeking in machine learning, psychology and neuroscience

For better or for worse, during our limited existence on earth, humans have altered the face of the world. We invented electricity, submarines, and airplanes, and developed farming and medicine to an extent that has massively changed our lives. There is little doubt that these extraordinary advances are made possible by our cognitive structure, particularly the ability to reason and build causal models of external events. In addition, we would argue that this extraordinary dynamism depends on our high degree curiosity, the burning desire to know and understand. Many animals, especially humans, seem to constantly seek knowledge and information in behaviors ranging from the very small (such as looking at a new storefront) to the very elaborate and sustained (such as reading a novel or carrying out research). Moreover, especially in humans, the search for information seems to be independent of a foreseeable profit, as if learning were reinforcing in and of itself.

Despite the importance of information-seeking for intelligent behavior, our understanding of its mechanisms is in its infancy. In psychology, research on curiosity surged during the 1960s and 1970s and subsequently waned [1] and has shown a moderate revival in neuroscience in recent years [2,3]. Our focus here is on evaluating three lines of investigation that are relevant to this question and have remained largely separate: studies of active learning and exploration in the machine learning and robotics fields, studies of eye movements in natural behavior, and studies of curiosity in psychology and neuroscience.

Glossary

Computational reinforcement learning: defines the problem of how to solve an MDP (or a POMDP) through learning (including trial and error), as well as associated computational methods.
Developmental robotics: research field modeling how embodied agents can acquire novel sensorimotor, cognitive, and social skills in an open-ended fashion over a developmental time span through integration of mechanisms that include maturation, intrinsically and extrinsically motivated learning, and self-organization.
Intrinsic and extrinsic rewards: normative accounts of behavior based on computational reinforcement learning and optimal control theory rely on the concept of a reward to assign value to alternative options, and often distinguish between extrinsic and intrinsic rewards. Extrinsic rewards are associated with classical task-directed learning and encode objectives such as finding food and winning a chess game. By contrast, intrinsic rewards are associated with internal cognitive variables such as esthetic pleasure, information-seeking, and epistemic disclosure. Intrinsic rewards may be based on uncertainty, surprise, and learning progress, and they may be either learnt or innate.
Markov decision process (MDP): defines the problem of selecting the optimal actions at each state to maximize future expected rewards in a Markov process.
Markov process (MP): mathematical model of the evolution of a system in which the prediction of a future state depends only on the current state and on the applied action, and not on the path by which the system reached the current state.
Metacognition: capability of a cognitive system to monitor its own abilities – for example, its knowledge, competence, memory, learning, or thoughts – and act according to the results of this monitoring. An example is a system capable of estimating how much confidence or uncertainty it has or how much learning progress it has achieved, and then using these estimates to select actions.
Optimization: mechanism that is often used in machine learning to search for the best solution among competing solutions with regard to given criteria. Stochastic optimization is an approach in which improvements over current best estimates of the solution are searched by iteratively trying random variations of these best estimates.
POMDP: extension of MDP for the case where the state is not entirely or directly observable but is described by probability distributions.

Corresponding author: Gottlieb, J. (jg2141@columbia.edu).

1364-6613/$ – see front matter
 http://dx.doi.org/10.1016/j.tics.2013.09.001

As described below, although they use different terminology and methods, these three lines of research grapple with strikingly similar questions and propose overlapping mechanisms. We suggest that achieving closer integration holds much promise for expansion of this research field.

Information-seeking obeys the imperative to reduce uncertainty and can be extrinsically or intrinsically motivated

Multiple paradigms have been devoted to the study of exploration and have used a common definition of this process as the choice of actions with the goal of obtaining information. Although exploratory actions can involve physical acts, they are distinct from other motor acts in that their primary goal is not to exert force on the world, but to alter the observer's epistemic state. For instance, when we turn to look at a new storefront, the goal of the orienting action is not to affect a change in the external world (as we would, for instance, when we reach for and grasp an apple). Instead, the goal is to obtain information. Thus, the key questions we have to address when studying exploration and information-seeking pertain to the ways in which observers handle their own epistemic states, and specifically, how observers estimate their own uncertainty and find strategies that reduce that uncertainty.

Theoretical and empirical considerations show that the motivations behind this process can be diverse and derive from extrinsic or intrinsic factors. In extrinsically motivated contexts, information-gathering is a means to an end: it is used to maximize the agent's progress toward a separate goal. Paradigmatic examples of this type of sampling are the eye movements that subjects make in natural behavior, such as glancing at the traffic light at a busy intersection [4], an example we discuss in detail below (Figure 1A). In reinforcement learning (RL) terms, such task-related information sampling is a feature of pure exploitation. The agent is engaged in a task that seeks to maximize an extrinsic reward (e.g., food or money) and information-gathering is an intermediate step in attaining this reward. A more complex form of this process arises while learning a task, when an agent wishes to reach a goal but must explore to discover an appropriate strategy for reaching that goal (e.g., learning how to drive or how to play chess).

In contrast with such task-related sampling, information-seeking can also be intrinsically motivated, that is, a goal in and of itself. The fact that animals, and particularly

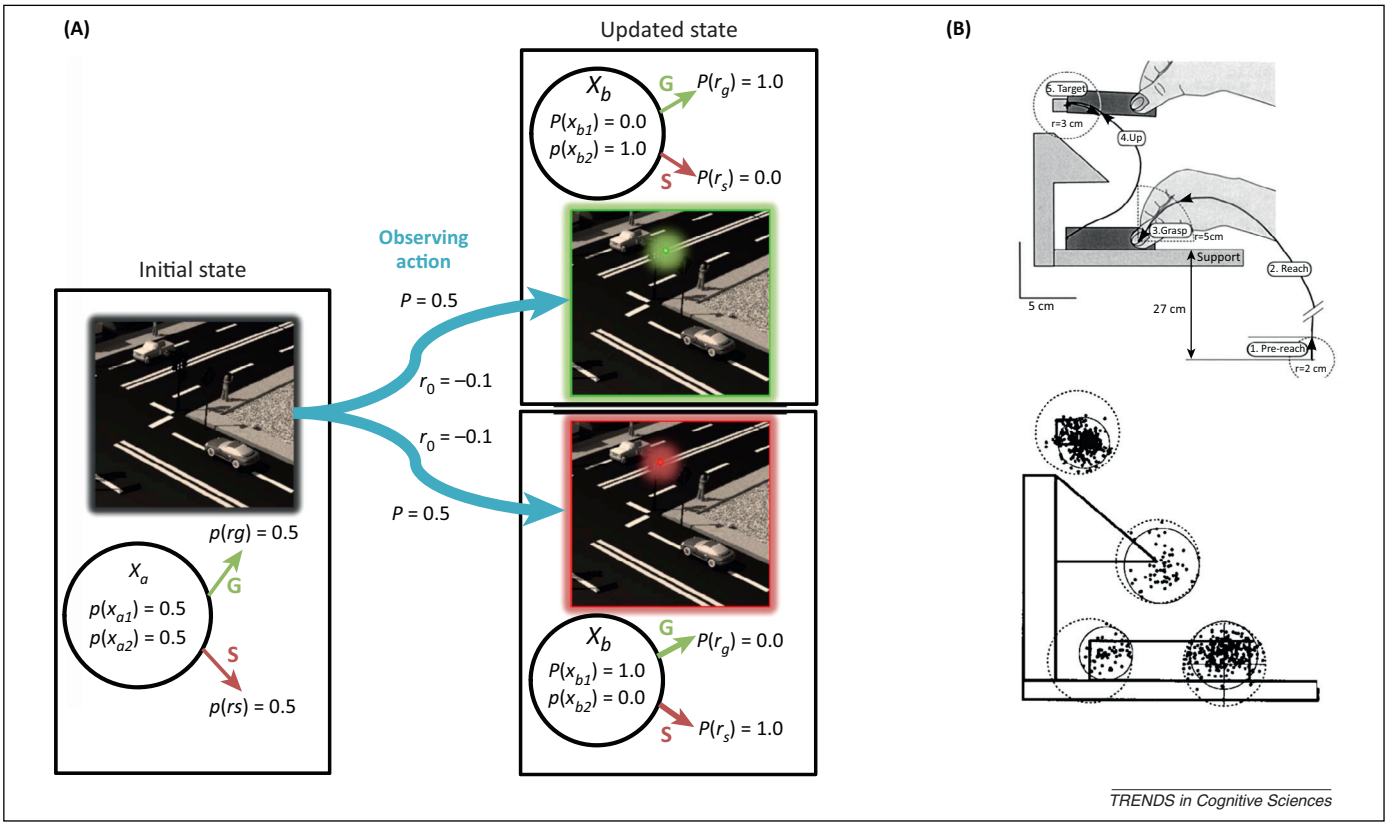

Figure 1. Information searches while executing a known task. **(A)** Description of an observing action – looking at the traffic light at an intersection – using a partly observable Markov decision process (POMDP). The observer starts in state x_a, where he arrives at an intersection and can take two actions, stop (S) or go (G). State x_a can be described as a stochastic mixture of two states, x_{a1} and x_{a2}, which are consistent with stopping and going, respectively, and have equal probabilities of 0.5. Thus, the expected probability of successfully crossing the intersection for either action from this state is 0.5. (For simplicity we assume that the reward magnitudes are equal and constant for S and G.) Conversely, the agent can take an observing action that transitions him to state x_{b1} or x_{b2}. These two states are equiprobable ($P = 0.5$), and transitioning to each is associated with a cost, $r_o < 0$, related to the time and effort of the visual discrimination. However, these states no longer have uncertainty. The agent can take action S from x_{b1} (where the light is red) or action G from x_{b2} (where the light is green) and in either case have a high likelihood of success. **(B)** Eye movements during a visuomanual task. The top panel shows the manual task. Starting from an initial position (Pre-reach), subjects reach and grasp a rectangular block (Grasp), bring the block up to touch a target (Target), and return it to the initial position (not shown). The bottom panel shows the distribution of fixations during the task. Fixations precede the hand trajectory, with 90% of them (solid circles) falling within landmark zones (dotted circles), which are the task-relevant contact points (of the fingers with the block, the block with the table, and the block with the target) and a potential obstacle (the protruding corner). This fixation pattern is highly consistent across observers and, notably, includes almost no extraneous fixations or fixations on the hand itself. Adapted with permission from [27].

humans, show intrinsic curiosity and seem to avidly seek out information without an apparent ulterior motive suggests that the brain generates intrinsic rewards (see Glossary) that assign value to information, and raises complex questions regarding the computations underlying such rewards.

In the following sections we first discuss task-defined forms of information searches and their links to eye movements and attention, and then address the more complex curiosity-like mechanisms.

Task-directed searches for information through the prism of eye movements

Information sampling while executing a known task

Computational studies have shown that when an agent knows a task, the controllers implementing that task can select actions that harvest immediate rewards, or actions that have indirect benefits by facilitating future actions. For example, to get a cookie from a high shelf, an individual may first pull up a chair and climb on it before reaching and grasping the cookie. Information-gathering actions are a special type of intermediate step that obey the imperative to reduce uncertainty and adjudicate among competing interpretations. As we discuss below (Figure 1A), a driver who seeks to arrive home safely may glance at a traffic light before crossing an intersection as an intermediate step that reduces uncertainty and increases the chance of success of his future actions.

Many computational approaches can model this type of information-seeking in a sound way. A common one relies on partially observable Markov decision processes (POMDPs) [5,6] (alternative representations are presented in [7,8]). A POMDP is a mathematical formalism that describes a task as a series of states, each with a set of possible actions and immediate or future outcomes (rewards or punishments). The states are partially observable in the sense that their identities are not deterministic but described by probability distributions, making POMDPs useful tools for measuring uncertainty and the value of new information.

For illustration, we consider the task of driving safely across an intersection (Figure 1A). In a POMDP, the agent performing the task would be described as starting in an initial state (e.g., the intersection, denoted by x_a) from which he can choose two possible actions, S (stop) or G (go). However, the agent has uncertainty about the true nature of state x_a. For example, x_a may be a state for which only stopping receives a reward [$P(r_S) = 1$ and $P(r_G) = 0$] or for which only going receives a reward [$P(r_G) = 1$ and $P(r_S) = 0$]. If these two states are equally likely, the agent has maximal uncertainty and can only expect a rate of reward of 0.5 regardless of which action he takes. However, rather than acting directly under this uncertainty, the agent can choose to obtain more information through an intermediate observing action, such as looking at a traffic light. This action is modeled as a transition to a different state, x_b, for which the probability distributions are more clearly separated, and the agent can be certain whether the optimal action is to stop [the light is red and $P(r_S) = 1$, bottom panel] or to proceed [the light is green and $P(r_G) = 1$, top panel]. Regardless of which alternative is correct, the agent has a much higher likelihood of obtaining a reward after rather than before having taken the observing action.

It is clear from this POMDP-based analysis that the observing action is not valuable in and of itself, but only if it increases the likelihood of reward for subsequent actions. In the short term, the observing action delivers no reward but has a cost in terms of the time and effort needed to discriminate the information (indicated by $r_o < 0$ in Figure 1A). This cost becomes worthwhile only if the observing action transitions the agent to a better state, that is, if the cumulative future value of state x_b exceeds that of state x_a by a sufficient amount. Balancing the costs and benefits of information sampling can also be cast in an information theoretic perspective [9].

Whether or not information-sampling has positive value depends on two factors. First, the observer must know the significance of the information and use it to plan future actions. In the traffic example, glancing at the light is only valuable if the observer understands its significance and if he takes the appropriate action (e.g., if he steps on the brake at the red light). Thus, information value is not defined unless observers have prior knowledge of the task, a strong point to which we return below.

A second factor that determines information value is the observer's momentary uncertainty. Although uncertainty in a given task is typically associated with specific junctures that are learnt while learning the task (e.g., when driving we generally expect high uncertainty at an intersection) this may quickly change, depending on the observer's momentary state. If, for example, the driver looked ahead and saw that there was a car in the intersection, his uncertainty would be resolved at this point, rendering the light redundant and reducing the value of looking at it. This raises the question (which has not been explored so far in empirical investigations) to what extent informational actions such as task-related eye movements are habitual, immutable aspects of a task and to what extent they rapidly respond to changing epistemic conditions (Box 1).

Information-sampling while searching for a task strategy

Strategies for solving a task, including those for generating informational actions, are not known in advance and must also be learnt. This implies an exploratory process whereby the learner experiments, selects, and tries to improve alternative strategies. For instance, when learning how to drive, individuals must also learn where to look to efficiently sample information; when learning chess, players must discover which strategy is most powerful in a given setting.

Deciding how to explore optimally when searching for strategy is a very difficult question, and is almost intractable in the general case. This question has been tackled in machine learning for individual tasks as an optimization problem, in which the task is modeled as a cost function and the system searches for the strategy that minimizes this function. The search may use approaches ranging from reinforcement learning [5,10,11] to stochastic optimization [12], evolutionary techniques [13], and Bayesian optimization [14]. It may operate in model-based approaches by learning a model of world dynamics and using it to plan a solution [15], or it may directly optimize parameters of a

Box 1. Using eye movements to probe multiple processes of information-searching

Because of its amenability to empirical investigations and the large amount of research devoted to it, the oculomotor system is a potentially excellent model system for probing information-seeking. In human observers, eye movements show consistent patterns that are highly reproducible within and across observers, both in laboratory tasks and in natural behaviors [4,30]. Moreover, eye movements show distinctive patterns during learning versus skilled performance of visuomanual tasks [27], suggesting that they can be used to understand various types of information-searching.

In non-human primates, the main oculomotor pathways are well characterized at the level of single cells, and include sensory inputs from the visual system, and motor mechanisms mediated by the superior colliculus and brainstem motor nuclei that generate a saccade [75]. Interposed between the sensory and motor levels is an intermediate stage of target selection that highlights attention-worthy objects, and seems to encode a decision of when and to what to attend [32,76]. Importantly, responses to target selection are sensitive to expected reward in the lateral intraparietal area (LIP), the frontal eye field (FEF), the superior colliculus, and the substantia nigra pars reticulata [32,77–79], suggesting that they encode reinforcement mechanisms relevant for eye movement control.

Against the background of these results, the oculomotor system can be used to address multiple questions regarding exploration. Two especially timely questions pertain to saccade guidance by extrinsic and intrinsic rewards, and to the integration of various information-seeking mechanisms.

Multiple valuation processes select stimuli for eye movement control

Animal studies of the oculomotor system have so far focused on the coding of extrinsic rewards, using simple tasks in which monkeys receive juice for making a saccade. However, as we have discussed, eye movements in natural behavior are not motivated by physical rewards but by more indirect metrics related to the value of information. Although this question has not been systematically investigated, evidence suggests that such higher-order values may be encoded in target selection cells. One line of evidence shows the entity that is selected by cells is not the saccade itself but a stimulus of interest, and this selection is independent of extrinsic rewards that the monkeys receive for making a saccade [80,81]. A second line of evidence suggests that the cells reflect two reward mechanisms: direct associations between stimuli and rewards independent of actions, and a measure of, potentially, the information value of action-relevant cues (Figure 1A).

Evidence of the role of Pavlovian associations comes from a task in which monkeys were informed whether or not they would receive a reward by means of a visual cue. Importantly, the cues were not relevant for the subsequent action, that is, they did not allow the monkeys to plan ahead and increase their odds of success in the task. Nevertheless, positive (reward-predictive) cues had higher salience and elicited stronger LIP responses than negative (non-reward-predictive) cues [82]. This valuation differs fundamentally from the types of valuation we discussed in the text: not only is it independent of action, but it is also independent of uncertainty reduction, because the positive and negative cues provided equally reliable information about forthcoming rewards. Thus, the brain seems to employ a process that weights visual information based on direct reward associations, possibly related to a phenomenon dubbed 'attention for liking' in behavioral research [83]. Although a bias to attend to good news is suboptimal from a strict information-seeking perspective, it may be adaptive in natural behavior by rapidly drawing resources to potential rewards.

Additional evidence suggests that along with this direct stimulus–reward process, cells may be sensitive to an indirect (potentially normative) form of valuation such as that shown in Figure 1A. Thus, cells select cues that provide actionable information even when the monkeys examine those cues covertly, without making a saccade [84,85]. In addition, an explanation based on information value may explain a recent report that LIP neurons had enhanced responses for targets threatening large penalties in a choice paradigm [86]. Although this result is apparently at odds with the more commonly reported enhancement by appetitive rewards, in the task that the monkeys performed the high penalty target was also an informative cue. The monkeys were presented with choices between a high-penalty target and a rewarded or lower-penalty option, and in either case the optimal decision (which the monkeys took) was to avoid the former target and orient to the alternative options. It is possible, therefore, that LIP cells encode a two-stage process similar to that shown in Figure 1A, in which the brain first attends to the more informative high-penalty cue (without generating a saccade) and then, based on the information obtained from this cue, makes the final saccade to the alternative option.

In sum, existing evidence is consistent with the idea that target selection cells encode several valuation processes for selecting visual items. Understanding the details of these processes can shed much light on decisions guiding information seeking.

Integrating extrinsically and intrinsically motivated searches

Although information-sampling in task- and curiosity-driven contexts seems to answer a common imperative for uncertainty reduction, these behaviors evoke very different subjective experiences, suggesting that they recruit different mechanisms. The neural substrates of these differences are very poorly understood. Behavioral and neuropsychological studies in rats suggest that the brain contains two attentional systems. A system of 'attention for action' relies on the frontal lobe and directs resources to familiar and reliable cues, and a system of 'attention for learning' relies on the parietal lobe and preferentially weights novel, surprising, or uncertain cues [87,88]. However, this hypothesis has not been investigated in individual cells and it is not clear how it maps onto various information-seeking mechanisms. Thus, an important and open question concerns the representation of task-related versus open-ended curiosity mechanisms, and in particular the coding of factors such as the novelty, uncertainty, or surprise of visual cues. Although responses to novelty and uncertainty have been reported for cortical and subcortical structures [89], it is unknown how they relate to attention and eye movement control.

solution in model-free approaches [16]. Approximate general methods have been proposed in reinforcement learning that are based on random action selection, or give novelty or uncertainty bonuses (in addition to the task-specific reward) for collecting data in regions that have not been recently visited, or that have a high expected gain in information [10,15,17–21]; we discuss these factors in more detail below. Yet another approach to strategy-learning involves generalizing from previously learnt circumstances; for example, if I previously found food in a supermarket, I will look for a supermarket if I am hungry in a new town [22]. Many of these methods can be seen as a POMDP whose uncertainty does not apply to the task-relevant state but to the task parameters themselves. It is important to note that although these processes require significant exploration, they are goal-directed in the sense that they seek to maximize a separate, or extrinsic, reward (e.g., drive successfully to a destination).

Eye movements reflect active information searches

In foveate animals such as humans and monkeys, visual information is sampled by means of eye movements and in particular saccades, rapid eye movements that occur several times a second and point the fovea to targets of interest. Although some empirical and computational approaches have portrayed vision as starting with a passive input stage that simply registers the available information [23,24], the study of eye movements makes it clear

that information sampling is a highly active process [23,24]. Far from being a passive recipient, the brain actively selects and proactively seeks out the information it wishes to sample, and it has been argued that this active process plays a key role in the construction of conscious perception [25].

Converging evidence suggests that when deployed in the service of a task, eye movements may be explained by the simple imperative to sample information to reduce uncertainty regarding future states [4,26]. In well-practiced tasks that involve visuomanual coordination (such as moving an object to a target point), the eyes move ahead of the hand to critical locations such as potential collision or target points, and wait there until the hand has cleared those locations (Figure 1B) [27]. Notably, the eyes never track the hand, which relies on motor and proprioceptive guidance and, for short periods of time, follows a predictable path; instead, they are strategically deployed to acquire new information. Additional evidence that gaze is proactively guided by estimates of uncertainty comes from a virtual reality study in which groups of observers walked along a track [28]. Subjects preferentially deployed gaze to oncoming pedestrians whose trajectory was expected to be uncertain (i.e., who had a history of veering onto a collision course) relative to those who had never shown such deviations. This suggests that observers monitor the uncertainty or predictability of external items and use these quantities proactively to deploy gaze (i.e., before and regardless of an actual collision). Finally, the eye movement patterns made while acquiring a task differ greatly from those made after learning [29,30], suggesting that eye movements are also coupled to exploration for a task strategy. These observations, together with the fact that eye movements are well investigated at the single neuron level in experimental animals and use value-based decision mechanisms [31,32], suggest that the oculomotor system may be an excellent model system for understanding information-seeking in the context of a task (Box 1).

Curiosity and autonomous exploration

Whereas in the examples discussed so far the goal of the task is known in advance and can be quantified in terms of extrinsic rewards, the open-ended nature of curiosity-like behaviors raises more difficult questions. To explain such behaviors and the high degree of motivation associated with them, it seems necessary to assume that the brain generates intrinsic rewards that assign value to learning or information *per se* [33]. Some support for this idea comes from the observation that the dopaminergic system, the chief reward system of the brain, is sensitive to intrinsic rewards [34], responds to anticipated information about rewards in monkeys [35], and is activated by paradigms that induce curiosity in humans [2,3]. However, important questions remain regarding the nature of intrinsic rewards at what David Marr would call the computational, representational, and physical levels of description [36]. At the computational level, it is not clear why the brain should generate intrinsic motivation for learning, how such motivation would benefit the organism, and the problems that it seeks to resolve. At the algorithmic and physical levels, it is unclear how these rewards are calculated and how they are implemented in the brain. We discuss each question in turn.

The benefits and challenges of information-seeking

The most likely answer to why intrinsic motivation for learning is generated is that such a motivation maximizes long-term evolutionary fitness in rapidly changing environmental conditions (e.g., due to human social and cultural structures, which can evolve much faster than the phylogenetic scale). For example, Singh *et al.* used computer simulations to show that, in dynamic contexts, even if the objective fitness/reward function of an organism is to survive and reproduce, it may be more efficient to evolve a control architecture that encodes an innate surrogate reward function rewarding learning *per se* [37]. The benefits of such a system arise because of the limited cognitive capacities of the agent (i.e., inability to solve the fitness function directly) [38–40] or because information or skills that are not immediately useful may be reused in the future. This idea resonates with the free-energy principle, which states that the possession of a large array of skills can be useful in avoiding future surprises by 'anticipating a changing and itinerant world' [41,42]. In fact, it is possible to show that making the environment predictable (by minimizing the dispersion of its hidden states) necessarily entails actions that decrease uncertainty about future states [42]. This idea resonates with the notion of Gestalt psychologists that humans have a 'need for cognition', that is, an instinctive drive to make sense of external events that operates automatically in mental processes ranging from visual segmentation to explicit inference and causal reasoning [1]. In one way or another, all these formulations suggest that information-seeking, like other cognitive activities, acquire value through long-term evolutionary selection in dynamic conditions.

If we accept that learning for its own sake is evolutionarily adaptive, a question arises regarding the challenges that such a system must solve. To appreciate the full scope of this question, consider the challenges that are faced by a child who learns life skills through an extended period of play and exploration [43–47]. One salient fact regarding this process is the sheer vastness of the learning space, especially given the scarce time and energy available for learning. In the sensorimotor domain alone, and in spite of significant innate constraints, a child needs to learn to generate an enormous repertoire of movements by orchestrating multiple interdependent muscles and joints that can be accessed at many hierarchical levels and interact in a potentially infinite number of ways with a vast number of physical objects/situations [48]. At the same time, in the cognitive domain, infants must acquire a vast amount of factual knowledge, rules, and social skills.

A second salient fact regarding this question is that while sampling this very large space, a child must avoid becoming trapped in unlearnable situations in which he cannot detect regularities or improve. In stark contrast to controlled laboratory conditions in which subjects are given solvable tasks, in real world environments many of the activities that an agent can choose for itself are inherently unlearnable either because of the learner's own limitations or because of irreducible uncertainty in the

problem itself. For instance, a child is bound to fail if she tries to learn to run before learning to stand, or tries to predict the details of the white noise pattern on a television screen. Thus, the challenge of an information-seeking mechanism is to efficiently learn a large repertoire of diverse skills given limited resources, and avoid being trapped in unlearnable situations.

A number of processes for open-ended exploration have been described in the literature that, as we describe below, have individual strengths and weaknesses and may act in complementary fashion to accomplish these goals. We consider first heuristics based on random action selection, novelty, or surprise, followed by deliberate strategies for acquiring knowledge and skills.

Randomness, novelty, surprise and uncertainty

In neuroscience research, the most commonly considered exploration strategies are based on random action selection or automatic biases toward novel, surprising or uncertain events. Sensory novelty, defined as a small number of stimulus exposures, is known to enhance neural responses throughout visual, frontal, and temporal areas [49] and activate reward-responsive dopaminergic areas. This is consistent with the theoretical notion that novelty acts as an intrinsic reward for actions and states that had not been recently explored or that produce high empirical prediction errors [50]. A more complex form of contextual novelty (also called surprise) has been suggested to account for attentional attraction toward salient events [51] and may be quantified using Bayesian inference as a difference between a prior and posterior world model [52] or as a high prediction error for high-confidence states [53]. Computational models have also incorporated uncertainty-based strategies, generating biases toward actions or states that have high variance or entropy [54,55].

As discussed above, actions driven by randomness, novelty, uncertainty, or surprise are valuable in allowing agents to discover new tasks. However, these actions have an important limitation in that they do not guarantee that an agent will learn. The mere fact that an event is novel or surprising does not guarantee that it contains regularities that are detectable, generalizable, or useful. Therefore, heuristics based on novelty can guide efficient learning in small and closed spaces when the number of tasks is small [56], but are very inefficient in large open ended spaces, where they only allow the agent to collect very sparse data and risk trapping him in unlearnable tasks [48,57,58]. This motivates the search for additional solutions that use more targeted mechanisms designed to maximize learning *per se*.

Information gap hypothesis of curiosity

On the basis of a synthesis of psychological studies on curiosity and motivation, Lowenstein proposed an information gap hypothesis to explain so-called specific epistemic curiosity, an observer's desire to learn about a specific topic [1]. According to the information gap theory, this type of curiosity arises because of a discrepancy between what the observer knows and what he would like to know, where knowledge can be measured using traditional measures of information. As a concrete illustration, consider a mystery novel in which the author initially introduces ten suspects who are equally likely to have committed a murder and the reader's goal is to identify the single true culprit. The reader can be described as wanting to move from a state of high entropy (or uncertainty, with 10 possible alternative murderers) to one of low entropy (with a single culprit identified), and his curiosity arises through his awareness of the difference between his current and goal (reference) uncertainty states. Defined in this way, curiosity can be viewed as a deprivation phenomenon that seeks to fill a need similar to other reference-point phenomena or biological drives. Just as animals seek to fill gaps in their physical resources (e.g., energy, sex, or wealth), they seek to fill gaps in their knowledge by taking learning-oriented actions. This brings us back to the imperative to minimize uncertainty about the state of the world, and suggests that this imperative is similar to a biological drive.

It is important to recognize, however, that whereas biological drives are prompted by salient and easily recognizable signs (e.g., somatic signals for hunger or sex), recognition and elimination of information gaps require a radically different, knowledge-based mechanism. First, the agent needs some prior knowledge to set the starting and the reference points. When reading a novel, we cannot estimate the starting level of entropy unless we have read the first few pages and acquired 'some' information about the setting. Similarly, we cannot set the reference point unless we know that mystery novels tell us about culprits, meaning that we should define our reference state in terms of the possible culprits rather than, for example, the properties of DNA. In other words, an agent cannot be curious outside a known context, similar to the requirements for prior knowledge that arise in extrinsically motivated eye movement control (Figure 1). Second, to define an information gap, an individual has to monitor her level of uncertainty, again similar to eye movement control.

Exploration based on learning progress (LP)

Despite its considerable strengths, a potential limitation of the information gap hypothesis is that agents may not be able to estimate the starting or desired levels of uncertainty given their necessarily limited knowledge of the broader context. In scientific research, for example, the results of an experiment typically open up new questions that were not foreseen, and it is not possible to estimate in advance the current entropy or the final desired state. Thus, a difficult question posed by this theory is how the brain can define information gaps in general situations.

An alternative mechanism for targeted learning has been proposed in the field of developmental robotics, and eschews this difficulty by tracking an agent's local learning progress without setting an absolute goal [48,57,58] following an early formulation presented by Schmidhuber [59]. The central objective of developmental robotics is to design agents that can explore in open-ended environments and develop autonomously without a pre-programmed trajectory, based on their intrinsic interest. A system that has been particularly successful in this regard explicitly measures the agent's learning progress in an activity (defined as an improvement in its predictions of the consequences of its actions [57] or in its ability to solve

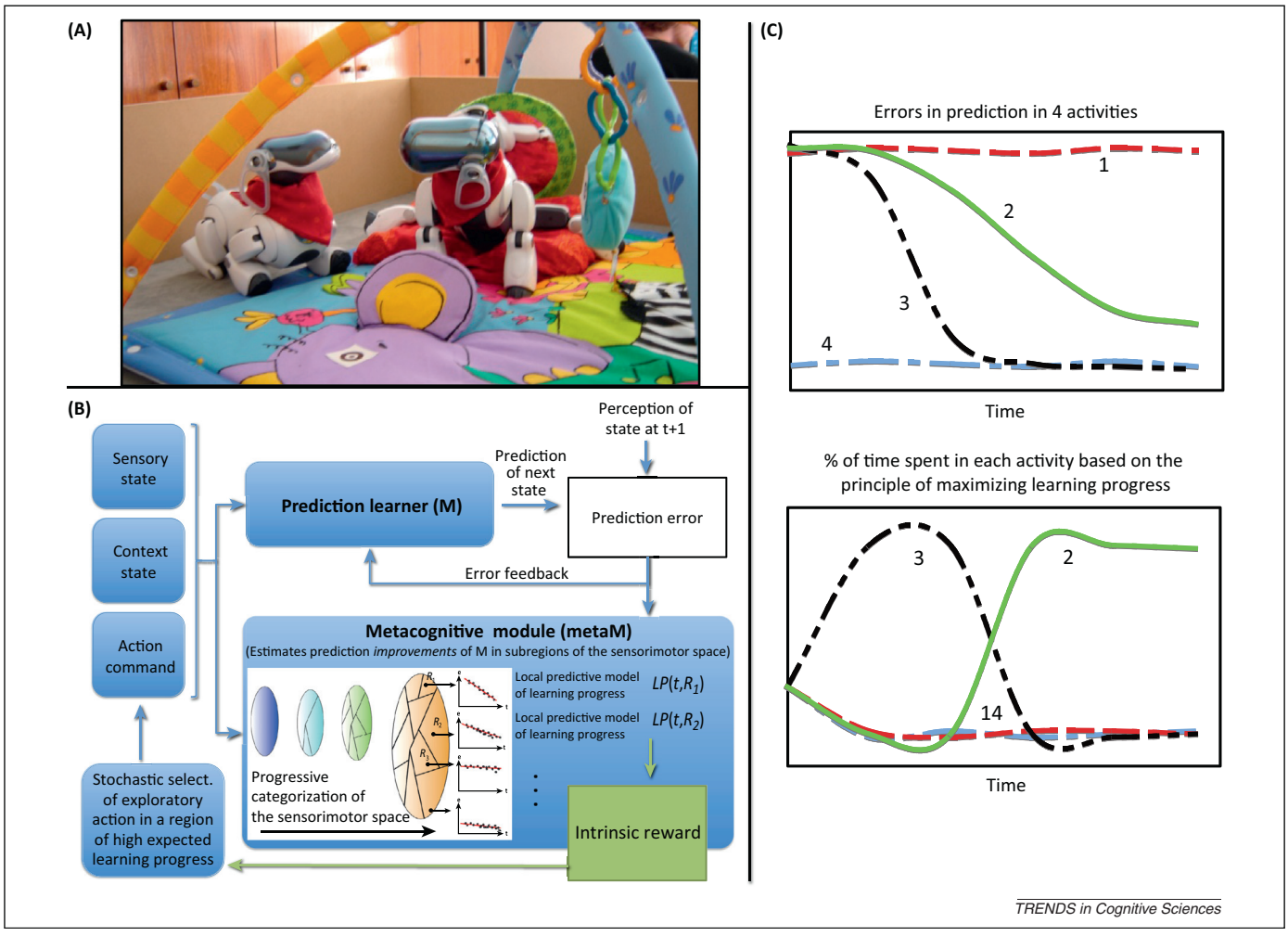

Figure 2. Curiosity-driven exploration through maximization of learning progress. **(A)** The Playground Experiment studies curiosity-driven exploration and how it can self-organize development, with a quadruped robot placed on an infant play mat with a set of nearby objects, as well as an 'adult' robot peer [57]. The robot is equipped with a repertoire of motor primitives parameterized by several continuous numbers, which can be combined to form a large continuous space of possible actions. The robot learns how to use and tune them to affect various aspects of its surrounding environment, and exploration is driven by maximization of learning progress. We observe the self-organization of structured developmental trajectories, whereby the robot explores objects and actions in a progressively more complex stage-like manner while acquiring autonomously diverse affordances and skills that can be reused later on. The robot also discovers primitive vocal interaction as a result of the same process [65,67]. Internally, the categorization system of such an architecture progressively builds abstractions that allow it to differentiate its own body (the self) from physical objects and animate objects (the other robot) [66]. **(B)** The R-IAC architecture implements this curiosity-driven process with several modules [48,57]. A prediction machine (M) learns to predict the consequences of actions taken by the robot in given sensory states. A meta-cognitive module (metaM) estimates the evolution of errors in prediction of M in various subregions of the sensorimotor space, which in turn is used to compute learning progress as an intrinsic reward. Because 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 regions, which incrementally model the creation and refining of activities/tasks. Then an action selection system chooses activities to explore for which estimated learning progress is high. This choice is stochastic in order to monitor other activities for which learning progress might increase, and is based on algorithms of the bandit family [46,90]. **(C)** Confronted with four sensorimotor activities characterized by different learning profiles (i.e., evolution of prediction errors), exploration driven by maximization of learning progress results in avoidance of activities already predictable (curve 4) or too difficult to learn to predict (curve 1) to focus first on the activity with the fastest learning rate (curve 3) and eventually, when the latter starts to reach a plateau, to switch to the second most promising learning situation (curve 2). This allows the creation of an organized exploratory strategy necessary to engage in open-ended development. Adapted with permission from [69].

self-generated problems over time [60,61]), and rewards activities in proportion to their ability to produce learning progress (Figure 2). Similar to an information-gap mechanism, this system produces a targeted search for information that drives the agent to learn. By using a local measure of learning, the system avoids difficulties associated with defining an absolute (and potentially unknowable) competence or epistemic goal.

This progress-based approach has been used most successfully in real-world situations. First, it allows robots to efficiently learn repertoires of skills in high dimensions and under strong time constraints while avoiding unfruitful activities that are either well learnt and trivial, or random and unlearnable [60,62–64]. Second, the system self-organizes development and learning trajectories that share fundamental qualitative properties with infant development, in particular the gradual shift of interest from simpler to more complex skills (Figure 2) [57,65–67]. This led to the hypothesis that some of the progressions in infant sensorimotor development may not be pre-programmed but emerge from the interaction of intrinsically motivated learning and the physical properties of the body and the environment [57,68,69]. Initially applied to sensorimotor tasks such as object manipulation, the approach also spontaneously led a robot to discover vocal communication with a peer (while traversing stages of babbling that resemble those of infants as a consequence of its drive to explore situations estimated to be learnable [65,67]).

In sum, a system based on learning progress holds promise for achieving efficient, intrinsically motivated exploration in large, open-ended spaces, as well as for self-organizing and ordering developmental stages. It must be noted, however, that although computationally powerful, this approach entails a complex meta-cognitive architecture for monitoring learning progress that still awaits empirical verification. Possible candidates for such a system include frontal systems that encode uncertainty or confidence in humans and monkeys [70–72] or respond selectively for behavioral change or the beginning of exploratory episodes [73,74]. However, a quantitative response to learning progress (which is distinct from phasic responses to novelty, surprise or arousal) has not been demonstrated in empirical investigations.

Concluding remarks

Although the question of active exploration is vast and cannot be exhaustively covered in a single review, we attempted to outline a few key ideas that are relevant to this topic from psychology, neuroscience, and machine learning fields. Three main themes emerge from the review. First, an understanding of information-seeking requires that we understand how agents monitor their own competence and epistemic states, and specifically how they estimate their uncertainty and generate strategies for reducing that uncertainty. Second, this question requires that we understand the nature of intrinsic rewards that motivate information-seeking and learning, and may impact cognitive development. Finally, eye movements are natural indicators of active information searching by the brain. By virtue of their amenability to neurophysiological investigations, the eyes may be an excellent model system for tackling this question, especially if studied in conjunction with computational approaches and the intrinsic reward and cognitive control mechanisms of the brain.

Acknowledgments

This work was supported by a Fulbright visiting scholar grant (A.B.), HSFP Cross-Disciplinary Fellowship LT000250 (A.B.), ERC Starting Grant EXPLORERS 240007 (PYO), and an Inria Neurocuriosity grant (J.G., P.Y.O., M.L., A.B.). We thank Andy Barto and two anonymous reviewers for their insightful comments on this paper.

References

1 Lowenstein, G. (1994) The psychology of curiosity: a review and reinterpretation. *Psychol. Bull.* 116, 75–98
2 Kang, M.J. *et al.* (2009) The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances memory. *Psychol. Sci.* 20, 963–973
3 Jepma, M. *et al.* (2012) Neural mechanisms underlying the induction and relief of perceptual curiosity. *Front. Behav. Neurosci.* 6, 5
4 Tatler, B.W. *et al.* (2011) Eye guidance in natural vision: reinterpreting salience. *J. Vis.* 11, 5–25
5 Kaelbling, L.P. *et al.* (1998) Planning and acting in partially observable stochastic domains. *Artif. Intell.* 101, 99–134
6 Dayan, P. and Daw, N.D. (2008) Decision theory, reinforcement learning, and the brain. *Cogn. Affect. Behav. Neurosci.* 8, 429–453
7 Bialek, W. *et al.* (2001) Predictability, complexity, and learning. *Neural Comput.* 13, 2409–2463
8 Singh, S. *et al.* (2004) Predictive state representations: a new theory for modeling dynamical systems. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 512–519, AUAI Press
9 Tishby, N. and Polani, D. (2011) Information theory of decisions and actions. In *Perception–Action Cycle* (Cutsuridis, V. *et al.*, eds), pp. 601–636, Springer
10 Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning: An Introduction*, MIT Press
11 Kober, J. and Peters, J. (2012) Reinforcement learning in robotics: a survey. In *Reinforcement Learning* (Wiering, W. and Van Otterlo, M., eds), In *Adaptation, Learning, and Optimization* (Vol. 12), pp. 579–610, Springer
12 Spall, J.C. (2005) *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, Wiley
13 Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Longman
14 Jones, D.R. *et al.* (1998) Efficient global optimization of expensive black-box functions. *J. Global Optim.* 13, 455–492
15 Brafman, R.I. and Tennenholtz, M. (2003) R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.* 3, 213–231
16 Sutton, R.S. *et al.* (2000) Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems* (Vol. 12), pp. 1057–1063, MIT Press
17 Sutton, R.S. (1990) Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the 7th International Conference on Machine Learning*, pp. 216–224, ICML
18 Dayan, P. and Sejnowski, T.J. (1996) Exploration bonuses and dual control. *Mach. Learn.* 25, 5–22
19 Kearns, M. and Singh, S. (2002) Near-optimal reinforcement learning in polynomial time. *Mach. Learn.* 49, 209–232
20 Kolter, J.Z. and Ng, A.Y. (2009) Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 513–520, ICML
21 Lopes, M. *et al.* (2012) Exploration in model-based reinforcement learning by empirically estimating learning progress. *Neural Inf. Process. Syst.* (NIPS 2012)
22 Dayan, P. (2013) Exploration from generalization mediated by multiple controllers. In *Intrinsically Motivated Learning in Natural and Artificial Systems* (Baldassarre, G. and Mirolli, M., eds), pp. 73–91, Springer
23 Blake, A. and Yuille, A.A.L. (1992) *Active Vision*, MIT Press
24 Tsotsos, J.K. (2011) *A Computational Perspective on Visual Attention*, MIT Press
25 O'Regan, J.K. (2011) *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*, Oxford University Press
26 Friston, K. and Ao, P. (2012) Free energy, value, and attractors. *Comput. Math. Methods Med.* 2012, 937860
27 Johansson, R.S. *et al.* (2001) Eye–hand coordination in object manipulation. *J. Neurosci.* 21, 6917–6932
28 Jovancevic-Misic, J. and Hayhoe, M. (2009) Adaptive gaze control in natural environments. *J. Neurosci.* 29, 6234–6238
29 Sailer, U. *et al.* (2005) Eye–hand coordination during learning of a novel visuomotor task. *J. Neurosci.* 25, 8833–8842
30 Land, M.F. (2006) Eye movements and the control of actions in everyday life. *Prog. Retin Eye Res.* 25, 296–324
31 Kable, J.W. and Glimcher, P.W. (2009) The neurobiology of decision: consensus and controversy. *Neuron* 63, 733–745
32 Gottlieb, J. (2012) Attention, learning, and the value of information. *Neuron* 76, 281–295
33 Berlyne, D. (1960) *Conflict, Arousal and Curiosity*, McGraw-Hill
34 Redgrave, P. *et al.* (2008) What is reinforced by phasic dopamine signals? *Brain Res. Rev.* 58, 322–339
35 Bromberg-Martin, E.S. and Hikosaka, O. (2009) Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron* 63, 119–126
36 Marr, D. (2010) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, MIT Press
37 Singh, S. *et al.* (2010) Intrinsically motivated reinforcement learning: an evolutionary perspective. *IEEE Trans. Auton. Ment. Dev.* 2, 70–82
38 Sorg, J. *et al.* (2010) Reward design via online gradient ascent. In *Advances in Neural Information Processing Systems* (Vol. 23), pp. 2190–2198, MIT Press
39 Lehman, J. and Stanley, K.O. (2011) Abandoning objectives: evolution through the search for novelty alone. *Evol. Comput.* 19, 189–223

40 Sequeira, P. *et al.* (2011) Emerging social awareness: exploring intrinsic motivation in multiagent learning. In *IEEE International Conference on Development and Learning (ICDL)*
41 Friston, K. (2010) The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138
42 Friston, K. *et al.* (2012) Free-energy minimization and the dark-room problem. *Front. Psychol.* 3, 130
43 Weng, J. *et al.* (2001) Autonomous mental development by robots and animals. *Science* 291, 599–600
44 Asada, M. *et al.* (2009) Cognitive developmental robotics: a survey. *IEEE Trans. Autonom. Ment. Dev.* 1, 12–34
45 Oudeyer, P.Y. (2010) On the impact of robotics in behavioral and cognitive sciences: from insect navigation to human cognitive development. *IEEE Trans. Auton. Ment. Dev.* 2, 2–16
46 Lopes, M. and Oudeyer, P.Y. (2012) The strategic student approach for life-long exploration and learning. In *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*
47 Baldassarre, G. and Mirolli, M. (2013) *Intrinsically Motivated Learning in Natural and Artificial Systems*, Springer-Verlag
48 Oudeyer, P.Y. *et al.* (2013) Intrinsically motivated learning of real-world sensorimotor skills with developmental constraints. In *Intrinsically Motivated Learning in Natural and Artificial Systems* (Baldassarre, G. and Mirolli, M., eds), pp. 303–365, Springer
49 Ranganath, C. and Rainer, G. (2003) Neural mechanisms for detecting and remembering novel events. *Nat. Rev. Neurosci.* 4, 193–202
50 Duzel, E. *et al.* (2010) Novelty-related motivation of anticipation and exploration by dopamine (NOMAD): implications for healthy aging. *Neurosci. Biobehav. Rev.* 34, 660–669
51 Boehnke, S.E. *et al.* (2011) Visual adaptation and novelty responses in the superior colliculus. *Eur. J. Neurosci.* 34, 766–779
52 Itti, L. and Baldi, P. (2009) Bayesian surprise attracts human attention. *Vis. Res.* 49, 1295–1306
53 Oudeyer, P.Y. and Kaplan, F. (2007) What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobot.* 1, 6
54 Cohn, D.A. (1996) Active learning with statistical models. *J. Artif. Intell. Res.* 4, 129–145
55 Rothkopf, C.A. and Ballard, D. (2010) Credit assignment in multiple goal embodied visuomotor behavior. *Front. Psychol.* 1, 173
56 Thrun, S. (1995) Exploration in active learning. In *Handbook of Brain Science and Neural Networks* (Arbib, M.A., ed.), pp. 381–384, MIT Press
57 Oudeyer, P.Y. *et al.* (2007) Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286
58 Schmidhuber, J. (2013) Maximizing fun by creating data with easily reducible subjective complexity. In *Intrinsically Motivated Learning in Natural and Artificial Systems* (Baldassarre, G. and Mirolli, M., eds), pp. 95–128, Springer
59 Schmidhuber, J. (1991) Curious model-building control systems. In *Proceedings of the International Joint Conference on Neural Networks*, pp. 1458–1463, IEEE
60 Baranes, A. and Oudeyer, P.Y. (2013) Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robot. Auton. Syst.* 61, 49–73
61 Srivastava, R.K. *et al.* (2013) First experiments with PowerPlay. *Neural Netw.* 41, 130–136
62 Pape, L. *et al.* (2012) Learning tactile skills through curious exploration. *Front. Neurorobot.* 6, 6
63 Ngo, H. *et al.* (2012) Learning skills from play: artificial curiosity on a Katana robot arm. In *Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE
64 Nguyen, S.M. and Oudeyer, P.Y. (2013) Socially guided intrinsic motivation for robot learning of motor skills. *Auton Robots* http://dx.doi.org/10.1007/s10514-013-9339-y
65 Oudeyer, P.Y. and Kaplan, F. (2006) Discovering communication. *Connect. Sci.* 18, 189–206
66 Kaplan, F. and Oudeyer, P.Y. (2011) From hardware and software to kernels and envelopes: a concept shift for robotics, developmental psychology and brain science. In *Neuromorphic and Brain-based Robots* (Krichmar, J.L. and Wagatsuma, H., eds), pp. 217–250, Cambridge University Press
67 Moulin-Frier, C. and Oudeyer, P.Y. (2012) Curiosity-driven phonetic learning. In *Proceedings of the 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pp. 1–8, IEEE
68 Smith, L.B. and Thelen, E. (2003) Development as a dynamic system. *Trends Cogn. Sci.* 7, 343–348
69 Kaplan, F. and Oudeyer, P.Y. (2007) In search of the neural circuits of intrinsic motivation. *Front. Neurosci.* 1, 225–236
70 Fleming, S.M. *et al.* (2010) Relating introspective accuracy to individual differences in brain structure. *Science* 329, 1541–1543
71 Fleming, S.M. *et al.* (2012) Prefrontal contributions to metacognition in perceptual decision making. *J. Neurosci.* 32, 6117–6125
72 De Martino, B. *et al.* (2013) Confidence in value-based choice. *Nat. Neurosci* 16, 105–110
73 Isoda, M. and Hikosaka, O. (2007) Switching from automatic to controlled action by monkey medial frontal cortex. *Nat Neurosci.* 10, 240–248
74 Quilodran, R. *et al.* (2008) Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron* 57, 314–325
75 Wurtz, R.H. and Goldberg, M.E., eds (1989) The Neurobiology of Saccadic Eye Movements. In *Reviews of Oculomotor Research* (Vol. III), Elsevier
76 Thompson, K.G. and Bichot, N.P. (2005) A visual salience map in the primate frontal eye field. *Prog. Brain Res.* 147, 251–262
77 Ding, L. and Hikosaka, O. (2006) Comparison of reward modulation in the frontal eye field and caudate of the macaque. *J. Neurosci.* 26, 6695–6703
78 Isoda, M. and Hikosaka, O. (2008) A neural correlate of motivational conflict in the superior colliculus of the macaque. *J. Neurophysiol.* 100, 1332–1342
79 Yasuda, M. *et al.* (2012) Robust representation of stable object values in the oculomotor basal ganglia. *J. Neurosci.* 32, 16917–16932
80 Gottlieb, J. and Balan, P.F. (2010) Attention as a decision in information space. *Trends Cogn. Sci.* 14, 240–248
81 Suzuki, M. and Gottlieb, J. (2013) Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nat. Neurosci.* 16, 98–104
82 Peck, C.J. *et al.* (2009) Reward modulates attention independently of action value in posterior parietal cortex. *J Neurosci* 29, 11182–11191
83 Hogarth, L. *et al.* (2010) Selective attention to conditioned stimuli in human discrimination learning: untangling the effects of outcome prediction, valence, arousal and uncertainty. In *Attention and Associative Learning: From Brain to Behaviour* (Mitchell, C.J. and Le Pelley, M.E., eds), pp. 71–98, Oxford University Press
84 Thompson, K.G. *et al.* (2005) Neuronal basis of covert spatial attention in the frontal eye field. *J. Neurosci.* 25, 9479–9487
85 Oristaglio, J. *et al.* (2006) Integration of visuospatial and effector information during symbolically cued limb movements in monkey lateral intraparietal area. *J. Neurosci.* 26, 8310–8319
86 Leathers, M.L. and Olson, C.R. (2012) In monkeys making value-based decisions, LIP neurons encode cue salience and not action value. *Science* 338, 132–135
87 Holland, P.C. and Maddux, J-M. (2010) Brain systems of attention in associative learning. In *Attention and Associative Learning: From Brain to Behaviour* (Mitchell, C.J. and Le Pelley, M.E., eds), pp. 305–350, Oxford University Press
88 Pearce, J.M. and Mackintosh, N.J. (2010) *Two Theories of Attention: A Review and a Possible Integration*, Oxford University Press
89 Bach, D.R. and Dolan, R.J. (2012) Knowing how much you don't know: a neural organization of uncertainty estimates. *Nat. Rev. Neurosci.* 13, 572–586
90 Bubeck, S. and Cesa-Bianchi, N. (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.* 5, 1–122