

Published in final edited form as:

Trends Cogn Sci. 2013 November ; 17(11): 585–593. doi:10.1016/j.tics.2013.09.001.

Information seeking, curiosity and attention: computational and neural mechanisms

Jacqueline Gottlieb^{1,2}, Pierre-Yves Oudeyer^{3,4}, Manuel Lopes^{3,4}, and Adrien Baranes¹

¹Department of Neuroscience, Columbia University

²Kavli Institute for Brain Science, Columbia University

³Inria, France

⁴Ensta ParisTech, France

Summary

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 search for learning progress and information gain in curiosity-driven behavior. In addition, eye movements reflect visual information search in multiple conditions and are amenable for cellular-level investigations, suggesting 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 we humans have altered the face of the world. We invented electricity, submarines and airplanes, and we 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, this extraordinary dynamism is made possible by our high degree of curiosity - the burning *desire to know* and *understand*. Many animals, and especially humans, seem constantly to 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

© 2013 Elsevier Ltd. All rights reserved

Corresponding author: Jacqueline Gottlieb, PhD, Department of Neuroscience, Columbia University, 1051 Riverside Drive, Kolb Research Annex, Rm. 569, New York, NY 10032, Phone: 212-543-6931, ext. 500, Fax: 212-543-5816, jg2141@columbia.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Disclosure statement:

The authors declare that they have no conflict of interest.

out research). Moreover, especially in humans, the search for information can 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 (for a comprehensive review, see (Lowenstein 1994)), and shows a mild revival in neuroscience in recent years (Kang, Hsu et al. 2009; Jepma, Verdonchot et al. 2012). Our focus here is on evaluating recent developments related to this question from three lines of investigation that have remained largely separate – namely, 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. As we describe below, although using different terminology and methods, these three lines of research grapple with strikingly similar questions and propose overlapping mechanisms. We suggest that achieving a closer integration holds much promise for expanding this research field.

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

Multiple paradigms have been devoted to 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 and grasp an apple) but 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. Specifically, as we will see below, information seeking in its many manifestations is related to the ability of an observer to estimate his or her uncertainty and to find strategies that reduce that uncertainty.

While information seeking is often geared toward uncertainty reduction, 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 – i.e., 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 (Tatler, Hayhoe et al. 2011), an example we discuss in detail below (Fig. 1A). In reinforcement learning (RL) terms, task-related information sampling is in this case 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 the agent wishes to reach a goal but must explore in order to discover an appropriate strategy for reaching that goal (for instance, learning to drive or to play chess).

In *intrinsically motivated* contexts, on the other hand, the search for information is a goal in itself – a process we would intuitively call “curiosity” or “interest”. The fact that animals, and particularly humans, seem avidly to seek out information without an apparent ulterior motive suggests that the brain generates *intrinsic* rewards that assign value to information, and raises complex questions regarding the benefits and computations of such rewards.

In the following sections we first discuss task-defined forms of information search and their links with eye movements and attention, followed by the more complex curiosity-like mechanisms.

Task-directed search 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, in order to get a cookie from a high shelf, you 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 (Fig. 1A) a driver who seeks to arrive safely home 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, and a commonly used one relies on partially-observable Markov decision processes (POMDPs) (Kaelbling, Littman et al. 1998; Dayan and Daw 2008) (see (Bialek, Nemenman et al. 2001; Singh, James et al. 2004) for alternative representations). A POMDP is a mathematical formalism that describes a task as a series of states, each with its 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 purposes of illustration, let us consider the example we mentioned above – the task of driving safely across an intersection (Fig. 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 where only stopping receives a reward ($p(r_S) = 1$ and $p(r_G) = 0$) or one where 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 0.5 rate of reward no matter 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 , where 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 turns out to be 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, as an intermediate step in a sequence, an observing action is only valuable if it increases the likelihood of reward of *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 Fig. 1A). This cost becomes worthwhile only if the observing action transitions the agent to a better state - i.e., 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 (Tishby and Polani 2011).

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). This makes the strong computational point that uncertainty and information value are not defined unless observers have prior knowledge of the links between stimuli and actions.

A second factor that determines information value is the observer's momentary uncertainty. Although in a given task uncertainty is typically associated with specific junctures (e.g., when driving one generally expects 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 so far explored 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 but 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 one must also learn where to look to efficiently sample information, and when learning chess one 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, where 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 (Kaelbling, Littman et al. 1998; Sutton and Barto 1998; Jens Kober 2012), to stochastic optimization (Spall 2005), evolutionary techniques (Goldberg 1989), or Bayesian optimization (Jones, Schonlau et al. 1998). It may operate in model-based approaches, by learning a model of the world

dynamics and using it to plan a solution (Brafman and Tenenbholz 2003), or it may directly optimize parameters of a solution in model-free approaches (Sutton, McAllester et al. 1999). Approximate *general* methods have been proposed in reinforcement learning, which 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 (Sutton ; Dayan and Sejnowski 1996; Sutton and Barto 1998; Kearns and Singh 2002; Brafman and Tenenbholz 2003; Kolter and Ng; Lopes, Lang et al.) (we discuss these factors in more detail in the following section). Yet another approach to strategy learning involves generalizing from previously learnt circumstances (e.g., if I previously found food in a supermarket, I will look for a supermarket if I am hungry in a new town) (Dayan 2013). Many of these methods can be seen as a POMDP whose uncertainty is not on the task-relevant state but on the task parameters themselves. It is important to note that, while 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 search

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. While some empirical and computational approaches have portrayed vision as starting with a *passive* input stage that simply registers the available information (Blake and Yuille 1992; Tsotsos 2011), the active control of eye movements makes it clear that information sampling is a highly active process (Blake and Yuille 1992; Tsotsos 2011). Far from being a passive recipient, the brain actively *selects* and proactively *seeks out* the information it wishes to sample, and this active process has been argued to play a key role in the construction of conscious perception (O'Regan 2011).

Converging evidence suggests that, when deployed in the service of a task, eye movements may be explained by the simple imperative to sample information in order to reduce uncertainty regarding future states (Tatler, Hayhoe et al. 2011; Friston and Ao 2012). In well-practiced tasks that involve visuo-manual 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 (Johansson, Westling et al. 2001) (Fig. 1B). 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 where groups of observers walked along a track (Jovancevic-Misic and Hayhoe 2009). 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 the 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 movements patterns made while acquiring a task differ greatly from those made after learning (Sailer, Flanagan et al. 2005; Land 2006), suggesting that eye movements are also coupled with exploring 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 (Kable and Glimcher 2009; Gottlieb 2012), 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

While for task-related behaviors the goal of a 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 related to learning or acquiring information (Berlyne 1960). Some support for this idea comes from the observation that the dopaminergic system, the brain's chief reward system, is sensitive to intrinsic rewards (Redgrave, Gurney et al. 2008), responds to anticipated information about rewards in monkeys (Bromberg-Martin and Hikosaka 2009) and is activated by paradigms that induce curiosity in humans (Kang, Hsu et al. 2009; Jepma, Verdonschot et al. 2012). However, important questions remain about the nature of intrinsic rewards at what David Marr would call the computational, representational and physical levels of description (Marr 2010). At the computational level, it is not clear *why* the brain should generate intrinsic motivation for learning - how such a motivation would benefit the organism and what are 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.

The benefits and challenges of information seeking

The most likely answer to the first, *why* question is that, by having an intrinsic motivation to learn, agents can maximize their 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). In such a context, Singh et al. (Singh, Lewis et al. 2010) have shown with computer simulations that 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*. The benefits of such a system arise because of the limited cognitive capacities of the agent (i.e., its inability to solve the fitness function directly) (Sorg, Lewis et al. 2010; Lehman and Stanley 2011; Sequeira, Melo et al. 2011) or because information or skills that are not immediately useful may be re-used 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 to avoid future surprises by “anticipating a changing and itinerant world” (Friston 2010; Friston, Thornton et al. 2012). In fact, it is possible to show that making the environment predictable (through minimizing the dispersion of its hidden states) necessarily entails actions that decrease uncertainty about future states (Friston, Thornton et al. 2012). This idea also resonates with the notion of Gestalt psychologists that humans have a “need for cognition” – i.e., 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 (Lowenstein 1994). In one way or another, all these formulations are consistent with the idea that intrinsically motivated cognitive activities, including information seeking, acquired value through long-term evolutionary selection in dynamic conditions.

If we accept that learning for its own sake is evolutionarily adaptive, the question arises regarding the challenges that such a system must solve. To appreciate the full scope of this question, let us consider the challenges that are faced by a child who learns life skills through an extended period of play and exploration (Weng, McClelland et al. 2001; Asada, Hosoda et al. 2009; Oudeyer 2010; Lopes and Oudeyer 2012; Baldassare and Mirolli 2013). One salient fact that emerges regarding this question 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, the 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 (Oudeyer, Baranes et al. 2013). Similarly 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, the child must avoid becoming trapped in *unlearnable* situations, i.e. where it cannot detect regularities or improve. In stark distinction with controlled laboratory conditions where subjects are given solvable tasks, in real world environments many of the activities that an agent can choose 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, and she is bound to fail if she tries to predict the details of white noise pattern on a television screen. Thus, the challenge of an information seeking mechanism is to learn efficiently a large repertoire of diverse skills given its limited resources, and avoid being trapped in unlearnable situations.

A number of processes for open-ended exploration have been described in the literature, which, 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/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 (Ranganath and Rainer 2003) and activate reward responsive dopamine-recipient 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 (Duzel, Bunzeck et al. 2010). A more complex form of "contextual novelty" (also called surprise) has been suggested to account for attentional attraction toward salient events (Boehnke, Berg et al.

2011) and may be quantified using Bayesian inference as a difference between a prior and posterior world model (Itti and Baldi 2009) or as a high prediction error for high-confidence states (Oudeyer and Kaplan 2007). Computational models have also incorporated uncertainty-based strategies, generating biases toward actions or states that have high variance or entropy (Cohn, Ghahramani et al. 1996; Rothkopf and Ballard 2010).

As discussed above, actions driven by randomness, novelty, uncertainty or surprise are valuable for 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, where the number of tasks is small (Thrun 1995) 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 (Oudeyer, Kaplan et al. 2007; Oudeyer, Baranes et al. 2013; Schmidhuber 2013). This motivates the search for additional solutions that use more targeted mechanisms designed to maximize learning *per se*.

Information gap hypothesis of curiosity

Based on a synthesis of psychological studies on curiosity and motivation, G.E. Lowenstein proposed an “information gap” hypothesis to explain so-called “specific epistemic curiosity” – an observer’s desire to learn about a specific topic (Lowenstein 1994). 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 with traditional measures of information. As a concrete illustration, consider a mystery novel where the author initially introduces 10 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 again 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, while biological drives are prompted by salient and easily recognizable signs (e.g., somatic signals for hunger or sex), recognizing and eliminating information gaps requires a radically different, knowledge-based mechanism. First, the agent needs some prior knowledge in order 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 rather than, for example, the properties of DNA (meaning that we should define our reference state in terms of the possible culprits). In other words, one cannot be curious about

what one does not know, similar to the requirements for prior knowledge of stimulus-action links that arise in eye movement control (Fig. 1). Second, to define an information gap one has to monitor one's level of uncertainty, again similar to eye movement control. Thus, the information gap theory links the study of epistemically-based action systems with that of motivated behaviors.

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 had not been foreseen, and it is not possible to estimate in advance what is the current entropy and 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, which eschews this difficulty by tracking an agent's local learning progress without setting an absolute goal (Oudeyer, Kaplan et al. 2007; Oudeyer, Baranes et al. 2013; Schmidhuber 2013) (following an early formulation presented in (Schmidhuber)). 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 (Oudeyer, Kaplan et al. 2007) or in its ability to solve self-generated problems over time (Baranes and Oudeyer 2013; Srivastava, Steunebrink et al. 2013)), and rewards activities in proportion to their ability to produce learning progress (see legend to Fig. 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 and to avoid unfruitful activities that are either well learnt and trivial, or which are random and unlearnable (L. Pape 2012; Ngo, Luciw et al. 2012; Baranes and Oudeyer 2013; Nguyen and Oudeyer 2013). 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 (Oudeyer and Kaplan 2006; Oudeyer, Kaplan et al. 2007; Kaplan and Oudeyer 2011; Moulin-Frier and Oudeyer 2012) (Fig. 2). 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 (Smith 2003; Kaplan and Oudeyer 2007; Oudeyer, Kaplan et al. 2007). Initially applied to sensorimotor tasks such as object manipulation, the approach was also shown to spontaneously lead 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 which the learner estimates to be learnable (Oudeyer and Kaplan 2006; Moulin-Frier and Oudeyer 2012)).

In sum, a system based on learning progress holds promise for achieving efficient, intrinsically motivated exploration in large open-ended spaces. It must be noted however that, while 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 the uncertainty or confidence in humans and monkeys (Fleming, Weil et al. 2010; Fleming, Huijgen et al. 2012; De Martino, Fleming et al. 2013) or which respond selectively for behavioral change or the beginning of exploratory episodes (Isoda and Hikosaka 2007; Quilodran, Rothe et al. 2008). 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.

Conclusions

While 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 the machine learning fields. Three main themes emerge from the review. First, understanding 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. Finally, eye movements are natural indicators of the brain's active information search. By virtue of their amenability to neurophysiological investigations, may be an excellent model system for tackling this question, especially if studied in conjunction with computational approaches and the brain's intrinsic reward and cognitive control mechanisms.

Acknowledgments

Fulbright visiting scholar grant (AB), HSFP Cross-Disciplinary Fellowship LT000250 (AB), ERC Starting Grant EXPLORERS 240007 (PYO), Inria Neurocuriosity grant (JG, PYO, ML, AB). We thank Andy Barto and two anonymous reviewers for their insightful comments on this paper.

References

- Asada M, Hosoda K, et al. Cognitive developmental robotics: A survey. *IEEE Trans. Autonomous Mental Development*. 2009; 1(1)
- Bach DR, Dolan RJ. Knowing how much you don't know: a neural organization of uncertainty estimates. *Nat Rev Neurosci*. 2012; 13(8):572–586. [PubMed: 22781958]
- Baldassare, G.; Mirolli, M. *Intrinsically motivated learning in natural and artificial systems*. Berlin: Springer-Verlag; 2013.
- Baranes A, Oudeyer PY. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*. 2013; 61(1):49–73.
- Berlyne, D. *Conflict, arousal and curiosity*. McGraw-Hill; 1960.

- Bialek W, Nemenman I, et al. Predictability, complexity, and learning. *Neural Computation*. 2001; 13(11):2409–2463. [PubMed: 11674845]
- Blake, A.; Yuille, AAL. *Active Vision*. Mit Press; 1992.
- Boehnke SE, Berg DJ, et al. Visual adaptation and novelty responses in the superior colliculus. *Eur J Neurosci*. 2011; 34(5):766–779. [PubMed: 21864319]
- Brafman RI, Tenenbalt M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*. 2003; 3:213–231.
- Bromberg-Martin ES, Hikosaka O. Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*. 2009; 63(1):119–126. [PubMed: 19607797]
- Cohn DA, Ghahramani Z, et al. Active learning with statistical models. *J Artificial Intelligence Research*. 1996; 4:129–145.
- Dayan, P. *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer; 2013. Exploration from generalization mediated by multiple controllers; p. 73-91.
- Dayan P, Daw ND. Decision theory, reinforcement learning, and the brain. *Cogn Affect Behav Neurosci*. 2008; 8(4):429–453. [PubMed: 19033240]
- Dayan P, Sejnowski TJ. Exploration bonuses and dual control. *Machine Learning*. 1996; 25(1):5–22.
- De Martino B, Fleming SM, et al. Confidence in value-based choice. *Nature Neuroscience*. 2013; 16(1):105–110.
- Ding L, Hikosaka O. Comparison of reward modulation in the frontal eye field and caudate of the macaque. *Journal of Neuroscience*. 2006; 26(25):6695–6703. [PubMed: 16793877]
- Duzel E, Bunzeck N, et al. Novelty-related motivation of anticipation and exploration by dopamine (NOMAD): implications for healthy aging. *Neurosci Biobehav Rev*. 2010; 34(5):660–669. [PubMed: 19715723]
- Fleming SM, Huijgen J, et al. Prefrontal contributions to metacognition in perceptual decision making. *The Journal of Neuroscience*. 2012; 32(18):6117–6125. [PubMed: 22553018]
- Fleming SM, Weil RS, et al. Relating introspective accuracy to individual differences in brain structure. *Science*. 2010; 329:1541–1543. [PubMed: 20847276]
- Friston K. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*. 2010; 11(2):127–138.
- Friston K, Ao P. Free energy, value, and attractors. *Comput Math Methods Med*. 2012; 2012:937860. [PubMed: 22229042]
- Friston K, Thornton C, et al. Free-energy minimization and the dark-room problem. *Frontiers in psychology*. 2012; 3
- Goldberg, DE. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley Longman Publishing Co; 1989.
- Gottlieb J. Attention, learning, and the value of information. *Neuron*. 2012; 76(2):281–295. [PubMed: 23083732]
- Gottlieb J, Balan PF. Attention as a decision in information space. *Trends in cognitive science*. 2010; 14(6):240–248.
- Hogarth, L.; Dickinson, A., et al., editors. *Attention and associative learning*. Oxford: Oxford University Press; 2010. Selective attention to conditioned stimuli in human discrimination learning: untangling the effects of outcome prediction, valence, arousal and uncertainty.
- Holland, PC.; Maddux, J-M. Brain systems of attention in associative learning. In: Mitchell, CJ.; Le Pelley, ME., editors. *Attention and associative learning*. Oxford University Press; 2010.
- Isoda M, Hikosaka O. Switching from automatic to controlled action by monkey medial frontal cortex. *Nat Neurosci*. 2007; 10(2):240–248. [PubMed: 17237780]
- Isoda M, Hikosaka O. A neural correlate of motivational conflict in the superior colliculus of the macaque. *J Neurophysiol*. 2008; 100(3):1332–1342. [PubMed: 18596188]
- Itti L, Baldi P. Bayesian surprise attracts human attention. *Vision research*. 2009; 49(10):1295–1306. [PubMed: 18834898]
- Jens Kober JP. Reinforcement learning in robotics: A survey. *Reinforcement Learning Adaptation, Learning, and Optimization*. 2012; 12:579–610.

- Jepma M, Verdonchot RG, et al. Neural mechanisms underlying the induction and relief of perceptual curiosity. *Front Behav Neurosci.* 2012; 6:5. [PubMed: 22347853]
- Johansson RS, Westling G, et al. Eye-hand coordination in object manipulation. *J Neurosci.* 2001; 21(17):6917–6932. [PubMed: 11517279]
- Jones DR, Schonlau M, et al. Efficient global optimization of expensive black-box functions. *Journal of Global optimization.* 1998; 13(4):455–492.
- Jovancevic-Misic J, Hayhoe M. Adaptive gaze control in natural environments. *J Neurosci.* 2009; 29(19):6234–6238. [PubMed: 19439601]
- Kable JW, Glimcher PW. The neurobiology of decision: consensus and controversy. *Neuron.* 2009; 63(6):733–745. [PubMed: 19778504]
- Kaelbling LP, Littman ML, et al. Planning and acting in partially observable stochastic domains. *Artificial intelligence.* 1998; 101(1):99–134.
- Kang MJ, Hsu M, et al. The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances memory. *Psychol Sci.* 2009; 20(8):963–973. [PubMed: 19619181]
- Kaplan F, Oudeyer P-Y. In search of the neural circuits of intrinsic motivation. *Frontiers in Neuroscience.* 2007; 1(1):225–225. [PubMed: 18982131]
- Kaplan, F.; Oudeyer, PY. *Neuromorphic and brain-based robots.* Krichmar, JL.; Wagatsuma, H., editors. 2011. p. 217-250.
- Kearns M, Singh S. Near-optimal reinforcement learning in polynomial time. *Machine Learning.* 2002; 49(2–3):209–232.
- Kolter, JZ.; Ng, AY. Near-Bayesian exploration in polynomial time; *Proceedings of the 26th Annual International Conference on Machine Learning*; 2009.
- Pape L, Controzzi CMOM, Cipriani C, Foerster A, Carrozza MC, Schmidhuber J. Learning tactile skills through curious exploration. *Frontiers in Neurorobotics.* 2012; 6(6)
- Land MF. Eye movements and the control of actions in everyday life. *Prog Retin Eye Res.* 2006; 25(3):296–324. [PubMed: 16516530]
- Leathers ML, Olson CR. In monkeys making value-based decisions, LIP neurons encode cue salience and not action value. *Science.* 2012; 338(6103):132–135. [PubMed: 23042897]
- Lehman J, Stanley KO. Abandoning objectives: evolution through the search for novelty alone. *Evolutionary Computation.* 2011; 19(2):189–223. [PubMed: 20868264]
- Lopes M, Lang T, et al. Exploration in model-based reinforcement learning by empirically estimating learning progress. *Neural Information Processing Systems (NIPS 2012).* 2012
- Lopes, M.; Oudeyer, P-Y. The strategic student approach for life-long exploration and learning; *Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on*; 2012.
- Lowenstein G. The psychology of curiosity: a review and reinterpretation. *Psychological Bulletin.* 1994; 116(1):75–98.
- Marr, D. *Vision: A computational investigation into the human representation and processing of visual information.* Cambridge, Massachusetts: The MIT Press; 2010.
- Moulin-Frier, C.; Oudeyer, P-Y. Curiosity-driven phonetic learning; *Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on*; 2012.
- Ngo, H.; Luciw, M., et al. Learning skills from play: Artificial curiosity on a Katana robot arm; *International Joint Conference on Neural Networks (IJCNN)*; 2012.
- Nguyen SM, Oudeyer P-Y. Socially guided intrinsic motivation for robot learning of motor skills. *Autonomous Robots.* 2013
- O'Regan, JK. *Oxford Scholarship*; 2011. Why red doesn't sound like a bell: Understanding the feel of consciousness.
- Oristaglio J, Schneider DM, et al. Integration of visuospatial and effector information during symbolically cued limb movements in monkey lateral intraparietal area. *J Neurosci.* 2006; 26(32): 8310–8319. [PubMed: 16899726]
- Oudeyer, P-Y.; Baranes, A., et al. *Intrinsically Motivated Learning in Natural and Artificial Systems.* Springer; 2013. Intrinsically motivated learning of real-world sensorimotor skills with developmental constraints; p. 303-365.

- Oudeyer PY. On the impact of robotics in behavioral and cognitive sciences: from insect navigation to human cognitive development. *IEEE Transactions on Autonomous Mental Development*. 2010; 2(1):2–16.
- Oudeyer PY, Kaplan F. Discovering communication. *Connection Science*. 2006; 18(2):189–206.
- Oudeyer PY, Kaplan F. What is Intrinsic Motivation? A Typology of Computational Approaches. *Frontiers of Neurorobotics*. 2007; 1(6–1):6.
- Oudeyer PY, Kaplan F, et al. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computations*. 2007; 11(2):265–286.
- Pearce, JM.; Mackintosh, NJ. Two theories of attention: a review and a possible integration. New York: Oxford University Press; 2010.
- Peck CJ, Jangraw DC, Suzuki M, et al. Reward modulates attention independently of action value in posterior parietal cortex. *J Neurosci*. 2009; 29(36):11182–11191. [PubMed: 19741125]
- Quilodran R, Rothe M, et al. Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron*. 2008; 57(2):314–325. [PubMed: 18215627]
- Ranganath C, Rainer G. Neural mechanisms for detecting and remembering novel events. *Nat Rev Neurosci*. 2003; 4(3):193–202. [PubMed: 12612632]
- Redgrave P, Gurney K, et al. What is reinforced by phasic dopamine signals? *Brain Res Rev*. 2008; 58(2):322–339. [PubMed: 18055018]
- Rothkopf CA, Ballard D. Credit assignment in multiple goal embodied visuomotor behavior. *Frontiers in Psychology*. 2010; 1(173)
- Sailer U, Flanagan JR, et al. Eye-hand coordination during learning of a novel visuomotor task. *J Neurosci*. 2005; 25(39):8833–8842. [PubMed: 16192373]
- Schmidhuber, J. Curious model-building control systems; *IEEE International Joint Conference on Neural Networks*; 1991.
- Schmidhuber, J. Maximizing fun by creating data with easily reducible subjective complexity. In: Baldassarre, G.; Mirolli, M., editors. *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer; 2013. p. 95–128.
- Sébastien Bubeck NC-B. Regret analysis of stochastic and nonstochastic multiarmed bandit problems. 2012
- Sequeira, P.; Melo, FS., et al. Emerging social awareness: Exploring intrinsic motivation in multiagent learning; *Development and Learning (ICDL), 2011 IEEE International Conference on*; 2011.
- Singh, S.; James, MR., et al. Predictive state representations: A new theory for modeling dynamical systems; *Proceedings of the 20th conference on Uncertainty in artificial intelligence*; 2004.
- Singh S, Lewis RL, et al. Intrinsically motivated reinforcement learning: An evolutionary perspective. *Autonomous Mental Development, IEEE Transactions on*. 2010; 2(2):70–82.
- Smith LB, Thelen E. Development as a dynamic system. *Trends in cognitive sciences*. 2003; 7(8):343–348. [PubMed: 12907229]
- Sorg, J.; Lewis, RL., et al. Reward design via online gradient ascent. *Advances in Neural Information Processing Systems (NIPS)*; 2010.
- Spall, JC. Introduction to stochastic search and optimization: estimation, simulation, and control. Wiley. com; 2005.
- Srivastava RK, Steunebrink BR, et al. First experiments with PowerPlay. *Neural Networks*. 2013; 41:130–136. [PubMed: 23465562]
- Sutton, RS. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming; *Proceedings of the seventh international conference (1990) on Machine learning*; 1990.
- Sutton, RS.; Barto, AG. *Reinforcement Learning: An Introduction*. MIT Press; 1998.
- Sutton RS, McAllester DA, et al. Policy gradient methods for reinforcement learning with function approximation. *Neural Information Processing Systems (NIPS)*. 1999; 99:1057–1063.
- Suzuki M, Gottlieb J. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nature Neuroscience*. 2013; 16(1):98–104.
- Tatler BW, Hayhoe MM, et al. Eye guidance in natural vision: reinterpreting salience. *J Vis*. 2011; 11(5):5. [PubMed: 21622729]

- Tatler BW, Hayhoe MN, et al. Eye guidance in natural vision: reinterpreting salience. *J Vis.* 2011; 11(5):5–25. [PubMed: 21622729]
- Thompson KG, Bichot NP. A visual salience map in the primate frontal eye field. *Prog Brain Res.* 2005; 147:251–262. [PubMed: 15581711]
- Thompson KG, Biscoe KL, et al. Neuronal basis of covert spatial attention in the frontal eye field. *J Neurosci.* 2005; 25(41):9479–9487. [PubMed: 16221858]
- Thrun, S. Exploration in active learning. In: Arbib, MA., editor. *Handbook of Brain Science and Neural Networks*. MIT Press; 1995. p. 381–384.
- Tishby, N.; Polani, D. Information theory of decisions and actions. In: Cutsuridis, V.; Hussain, A.; Taylor, JG., editors. *Perception-Action Cycle*. New York: Springer; 2011. p. 601–636.
- Tsotsos, JK. *A computational perspective on visual attention*. MIT Press; 2011.
- Weng J, McClelland J, et al. Autonomous mental development by robots and animals. *Science.* 2001; 291(5504):599–600. [PubMed: 11229402]
- Wurtz, RH.; Goldberg, ME. The neurobiology of saccadic eye movements, *Reviews of Oculomotor Research*. Vol. III. Amsterdam: Elsevier; 1989.
- Yasuda M, Yamamoto S, et al. Robust representation of stable object values in the oculomotor basal ganglia. *J Neurosci.* 2012; 32(47):16917–16932. [PubMed: 23175843]

Glossary

Developmental Robotics	Research field modeling how embodied agents can acquire novel sensorimotor, cognitive and social skills in 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 upon 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 like finding food or winning a chess game. In contrast, intrinsic rewards are associated with internal cognitive variables such as aesthetic pleasure, information seeking or epistemic disclosure. Examples of intrinsic rewards include measures of uncertainty, surprise or learning progress, and they may be either learnt or innate.
Markov process (MP)	mathematical model of the evolution of a system where the prediction of a future state depends only on the current state and on the applied action, but not on the path by which the system reached the current state.
Markov decision process (MDP)	defines the problem of selecting the optimal actions at each state in order to maximize future expected rewards.
POMDP	extension of MDP for the case where the state is not entirely or directly observable but is described by probability distributions.

**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.

Optimization

A 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 to optimization where improvements over current best estimates of the solution are searched by iteratively trying random variations of these best estimates.

Metacognition

The capability of a cognitive system to monitor its own abilities – e.g., its knowledge, competence, memory, learning or thoughts - and act based on 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 use these estimates to select actions.

Box 1: Using eye movements to probe multiple processes of information search

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 natural behaviors (Land 2006; Tatler, Hayhoe et al. 2011). Moreover, eye movements show distinctive patterns during the learning versus skilled performance of visuo-manual tasks (Sailer, Flanagan et al. 2005), suggesting that they can be used to understand various types of information search.

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 (Wurtz and Goldberg 1989). 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 (Thompson and Bichot 2005; Gottlieb 2012). 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 (Ding and Hikosaka 2006; Isoda and Hikosaka 2008; Gottlieb 2012; Yasuda, Yamamoto et al. 2012), suggesting that they encode reinforcement mechanisms relevant for eye movement control.

On 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 where 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. Evidence suggests (but does not yet conclusively establish) that such higher order values are encoded in target selection cells. Converging 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 (Gottlieb and Balan 2010; Suzuki and Gottlieb 2013). In addition, the cells seem to reflect two reward mechanisms – i.e., the learning of direct associations between stimuli and rewards independently of actions, and a measure of the information value of action-relevant cues (e.g., Fig. 1A).

Evidence for the role of Pavlovian associations comes from a task where monkeys were informed whether or not they will receive a reward by means of a visual cue. Importantly, the cues were not relevant for the subsequent action – i.e., did not allow the monkeys to plan ahead and increase their odds of success in the task. Nevertheless, the

positive (reward predictive) cues had higher salience and elicited stronger LIP responses than the negative (no-reward predicting) cues (Peck, Suzuki et al. 2009). 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, as 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 (Hogarth, Dickinson et al. 2010). 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, the cells may be sensitive to an indirect (potentially normative) form of valuation such as that shown in Fig. 1A. Thus, the cells select cues that provide actionable information even when the monkeys examine those cues covertly, without making a saccade (Thompson, Bischof et al. 2005; Oristaglio, Schneider et al. 2006). 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 (Leathers and Olson 2012). While 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 to the optimal decision (which the monkeys took) was to avoid the former target and orient to the alternative options. It is possible therefore that the LIP cells encoded a two-stage process similar to that shown in Fig. 1A, where the brain first attended to the more informative high penalty cue (without generating a saccade) and then, based on the information obtained from this cue, made 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, but the details of these processes remain poorly understood.

Integrating extrinsic and intrinsically motivated search

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” that relies on the frontal lobe and directs resources to familiar and reliable cues, and a system of “attention for learning” that relies on the parietal lobe and preferentially weights novel, surprising or uncertain cues (Holland and Maddux 2010; Pearce and Mackintosh 2010). However, this hypothesis has not been investigated in individual cells. Thus an important and wide 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. While

responses to novelty and uncertainty are reported in cortical and subcortical structures (Bach and Dolan 2012), it is unknown how they relate to attention and eye movement control.

Highlights

Information seeking can be driven by extrinsic or intrinsic rewards.

Curiosity may result from an intrinsic desire to reduce uncertainty.

Curiosity-driven learning is evolutionarily useful and can self-organize development.

Eye movements can provide an excellent model system for information seeking.

Computational and neural approaches converge to understand information seeking.

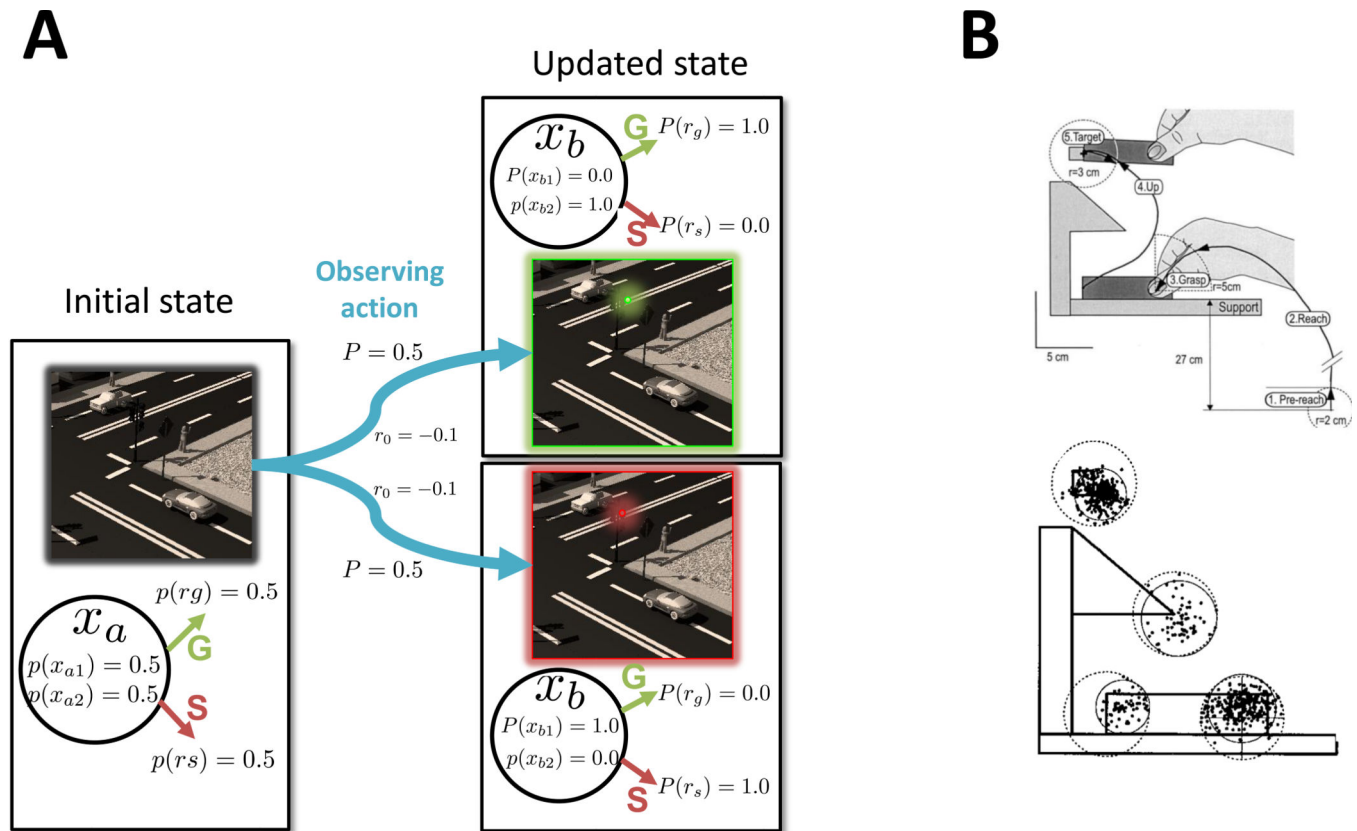


Fig. 1. Information search while executing a known task

A. Description of an observing action – looking at the traffic light at an intersection – using a POMDP. The observer starts in state x_a , where he arrives at an intersection and can take two actions, stop (S) or proceed (G). State x_a can be described as a stochastic mixture of two states x_{a1} and x_{a2} , which are consistent with, respectively, stopping or proceeding 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 reward magnitudes are equal and constant for S and G.) On the other hand, the agent can take an observing action that transitions him to states 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 reached and grasp a rectangular block (Grasp), brought the block up to touch a target (Target) and returned it to the initial position (not shown). The bottom panel shows the distribution of fixations during the task. Fixations precede the hand's 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 (Johansson, Westling et al. 2001).

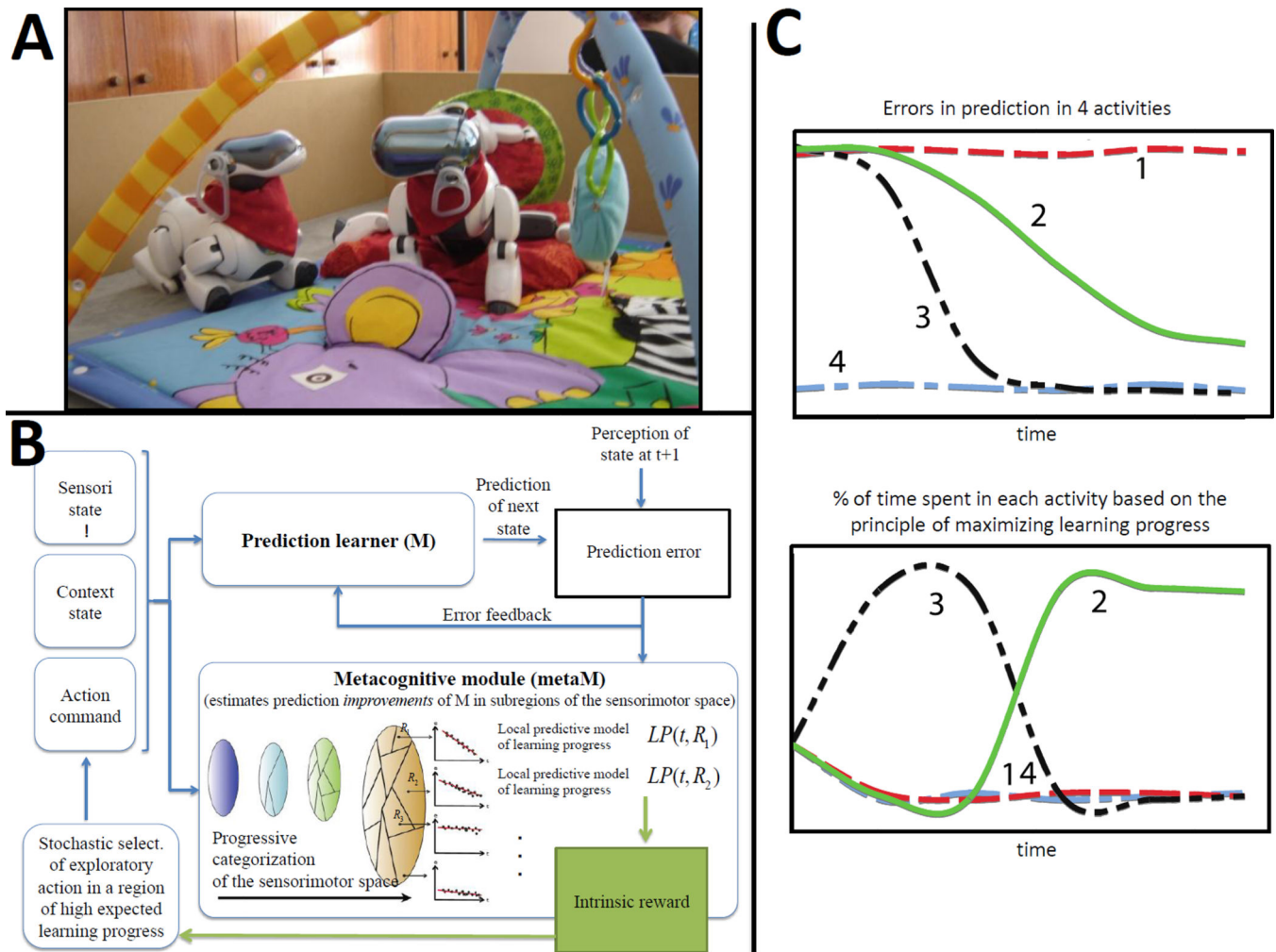


Fig. 2. Curiosity-driven exploration through maximization of learning progress

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. 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*. One observes the self-organization of structured developmental trajectories, where 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 (Oudeyer and Kaplan 2006; Moulin-Frier and Oudeyer 2012). Internally, the categorization system of such architecture progressively builds abstractions which allow it to differentiate its own body (the self) from physical objects but also animate objects (the other robot) (Kaplan and Oudeyer 2011). **B** The R-IAC architecture implements this curiosity-driven process with several modules (Oudeyer, Kaplan et al. 2007; Oudeyer, Baranes et al. 2013). 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. 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 regions, which incrementally model the creation and refining of activities/tasks. Then, an action selection system chooses activities to explore where estimated learning progress is high. This choice is stochastic in order to monitor other activities which learning progress might rise, and is based on algorithms of the bandit family (Lopes and Oudeyer ; Sébastien Bubeck 2012)

C. Confronted with four sensorimotor activities characterized by different learning profiles (i.e. evolution of prediction errors), exploration driven by maximizing learning progress results in avoiding activities already predictable (curve 4) or too difficult to learn to predict (curve 1), in order 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 (Kaplan and Oudeyer 2007).