Title

Information-Seeking, Learning and the Marginal Value Theorem:
A Normative Approach to Adaptive Exploration

Permalink

https://escholarship.org/uc/item/5339f64z

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 38(0)

Authors

Geana, Andra
Wilson, Robert C.
Daw, Nathaniel
et al.

Publication Date

2016

Peer reviewed

Information-Seeking, Learning and the Marginal Value Theorem: A Normative Approach to Adaptive Exploration

Andra Geana (ageana@princeton.edu)*+ Robert C. Wilson (bob@email.arizona.edu)°
Nathaniel Daw (ndaw@princeton.edu)*+ Jonathan D. Cohen (jdc@princeton.edu)*+

*Department of Psychology/Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540 USA
°Department of Psychology, University of Arizona, Tucson, AZ 85721

Abstract

Daily life often makes us decide between two goals: maximizing immediate rewards (exploitation) and learning about the environment so as to improve our options for future rewards (exploration). An adaptive organism therefore should place value on information independent of immediate reward, and affective states may signal such value (e.g., curiosity vs. boredom: Hill & Perkins, 1985; Eastwood et al. 2012). This tradeoff has been well studied in "bandit" tasks involving choice among a fixed number of options, but is equally pertinent in situations such as foraging, hunting, or job search, where one encounters a series of new options sequentially. Here, we augment the classic serial foraging scenario to more explicitly reward the development of knowledge. We develop a formal model that quantifies the value of information in this setting and how it should impact decision making, paralleling the treatment of reward by the marginal value theorem (MVT) in the foraging literature. We then present the results of an experiment designed to provide an initial test of this model, and discuss the implications of this information-foraging framework on boredom and task disengagement.

Keywords: exploration, explore-exploit tradeoff, information-seeking, decision making.

Introduction

All organisms face the frequent need to decide between persisting with one behavior (and the known rewards it brings), or switching to another. This tradeoff is well documented in the literature: stay-or-switch behavior has been studied in humans (Behrens et al., 2007) and non-human animals such as primates (Pearson et al. 2009), birds (Krebs, Kacelnik & Taylor 1978), rodents, and even non-vertebrates (Gallistel 1990) including the extent to which this follows optimal sampling strategies (Goldstone & Ashpole, 2004; Daw et al. 2006). When examining such decisions, it is helpful to distinguish between at least two types of circumstances under which an organism might choose to persist in or change its behavior: one involves situations in which rewards are which rewards are known up to stochasticity but either changing (as when foraging from a depleting patch) or varying in quality across options (as in encountering a series of candidate prey), so the decision whether to switch to other alternatives is a way to maximize current reward rate. The other involves situations in which the options' values are imperfectly known (as in bandit tasks) so that switching behavior may not yield immediate benefit, but may provide new information that can support learning and improvements in reward-rate over the longer term.

The first scenario has been extensively studied in the foraging literature. When choosing among behaviors with different reward opportunities (e.g. foraging patches) that are progressively depleting, or in circumstances in which (estimable) changes can happen outside the local environment, it is optimal to switch behavior when it is estimated that the value of the current behavior falls below the mean expected value of the available alternatives (Krebs & Inman, 1992). This policy can be shown to optimize immediate reward rate, as described by the marginal value theorem (MVT, Charnov, 1976), and numerous studies have found that animals' foraging behavior approximates this (Krebs, Kacelnik & Taylor 1978).

Most foraging scenarios of this type occur in environments with well-specified rewards, in which uncertainty usually stems from stable variance or hazard rates (risk), so it is possible to incorporate it into reward expectations through estimates of expected utility and the switching policy given by the MVT is optimal asymptotically (following all possible learning). However, in many circumstances reward opportunities may not only be stochastic, but the properties of this stochasticity may be unknown (Payzan-LeNestour et al., 2013). That is, uncertainty may stem from ambiguity rather than risk. In such circumstances, sampling the environment can provide information that, even if it is associated with immediate sacrifices in predictable reward rate, can be used to learn about the environment and improve reward rate over the longer term. We refer to such information-seeking as "exploration," to distinguish it from foraging choices that we define as the pursuit of alternative behaviors based on decisions involving reward opportunities with known distributional properties (e.g., mean and variance)[1].

The drive toward exploration has been well-documented in both human and animal literatures (e.g., Cohen, McClure & Yu, 2007), and there is rich evidence that under many scenarios, organisms will choose to sample the environment for useful information even at the cost of current or predictable reward (Wilson et al., 2014). This is the case, for instance, when we choose to try out the new special at a restaurant instead of sticking with our favorite dish, or when

[1] We make this qualitative distinction largely for the purpose of clarity, and to guide formal treatment, fully recognizing that real-world circumstances almost certainly fall along a continuum between these extremes and involve a mix of these two type of decisions.

we choose to watch a new show instead of rewatching an episode of an old favorite. In this way, exploration is different from foraging: though both involve the choice between persisting with the current action or switching away to something else, exploration is geared toward acquiring new information, while foraging is geared toward acquiring predictable sources of reward.

However, exploration of this sort has been studied largely in the context of “bandit” tasks – choice among a fixed set of options, whose properties must be learned from sampling – and not in the serial switching scenarios modeled by the MVT. Here, we propose a formal model for exploration that parallels the formulation of the MVT for reward, but applies it to maximizing information alongside reward. In particular, we augment the patch foraging scenario to more explicitly reward information gathering – this models, for instance, development of expertise when encountering a series of options, for instance, learning a trade over successive jobs or improving one’s dating skills – and study the behavior of optimal agents. We then present the results of an experiment designed to provide an initial test of this model, and discuss the implications of this information-foraging framework on boredom and task disengagement.

A Normative Model of Exploration

Paralleling circumstances to which the MVT has been applied, we model an environment consisting of local reward patches that offer different reward rates, with the model agent free to either stay within a patch to reap reward (exploit), or switch away to search for other patches, in this case with only partially or unknown characteristics (explore). Each patch is comprised of an environment in which the agent can earn rewards by making accurate predictions of the outcome of a stable stochastic process.

Upon "entering" a patch, the agent does not know the parameters of the stochastic process, but these can be learned by sampling. On each time step spent within a patch, the agent makes a prediction, and receives a reward proportional to the accuracy of the prediction. The longer an agent spends learning about a patch, the better its estimates of the underlying structure can become, and higher the reward it can receive. This distinguishes this task environment from the environment assumed in most studies of foraging: here, the patch becomes more rewarding with the passage of time (and sampling), rather than depleting.

An additional important assumption is that the properties of patches are not independent of one another, but rather reflect properties of a global environment from which they are drawn. Thus, within limits, sampling a local patch can provide information that is relevant to other patches. This is a property of many real-life environments, in which humans sequentially sampling different “patches” learn about the local structure while simultaneously learning about an overarching global structure (Diuk et al 2013). For instance, when going apple-picking, we learn about the quality and availability of fruit in each individual tree (so we could choose to move from a smaller, poorer tree to a better one), but at the same time we are also learning about the overall qualities of the orchard, so next time we go apple-picking me might choose an altogether different orchard.

Under this framework, exploiting a local patch obtained increasing local reward (fig. 1A), but exploring many local patches helps the agent learn the global structure faster, which would in turn allow it to make better choices earlier in the local patches. Depending on goals, therefore, it could be optimal to quit a local patch (even if it was yielding a high reward) and move to another patch, at the potential cost of a lower reward, for the sake of learning about the global environment that could improve returns in the future.

This local/global structure allowed us to model an environment with a distribution of available information paralleling the distribution of available reward in standard foraging environments. In other words, each patch held not only reward (which increased with time spent in patch), but also information (which decreased with time spent in patch). This generated a canonical explore-exploit decision tradeoff: maximize known rewards by staying within a patch (exploitation), or switch patches to acquire information (exploration). We constructed a model of this process, and used it to compare performance with a pure exploitation strategy, a strict foraging model, and human behavior in an empirical version of the task.

Model Assumptions

Each patch represented a stochastic environment in which the agent could earn rewards by making accurate predictions. Each patch had a hidden distribution with mean μ_i and standard deviation σ (which was the same between patches). On each time step spent inside the patch, the agent had to make a prediction relating to this distribution. The agent’s reward $r_{t,i}$ was proportional to the accuracy of the prediction (for a similar task design, see Nassar et al.’s (2010) “estimation task”), according to

$$r_{t,i} = \rho - PE_t, \tag{1}$$

where ρ represented the maximum amount of reward that an agent could earn (if its predictions were fully accurate), and PE_t represented the prediction error, computed as the difference between the agent’s prediction *Pr* and the actual number generated in the patch on time step *t*:

$$PE_t = Pr_t - N(\mu_i, \sigma) \tag{2}$$

Figure 1A shows the increase in reward with time spent in patch. An agent could spend as long as it wanted exploiting a patch, but each patch had a fixed chance of termination λ, meaning that on every time step the patch would end with probability λ, and continue with probability (1 – λ).

As explained in the previous section, all patches were connected under a higher-level, global structure. In other words, the underlying patch distribution parameter μ_i came from a global distribution with a (fixed) grand mean *M* and standard deviation *S*. Exploiting a local patch obtained increasing local reward (fig. 1A), but decreasing information (fig. 1B), while exploring more local patches helps the agent learn the global structure faster. There was also a global reward *R* associated with learning the global mean *M*. Our model tracked several quantities of interest as an agent exploited a patch with the above structure; for simplicity, we approximate the hierarchical estimation problem with nested error-driven updates. First, at each time step it computed an estimate of the local mean for patch *i* at time *t*, $\boldsymbol{\mu_{i,t}}$, as the average of all data points $x_{i,t}$ observed in that patch up to the current time:

$$\overline{\mu_{i,t}} = \frac{\sum_t x_{i,t}}{n} \qquad (3)$$

which can also be written in terms of the prediction error *PE* and a learning rate of 1/n, as

$$\overline{\mu_{i,t}} = \overline{\mu_{i,t-1}} + \frac{1}{n} * PE_t \qquad (4)$$

(Given the structure of the task, the optimal prediction at any time step was the current estimate of the mean, $\overline{\boldsymbol{\mu_{i,t}}}$, and our model assumed that the agent would always predict that mean). In addition to tracking the mean estimate for the patch, the model also used error-driven updating to estimate the variance of the local patch,

$$\sigma_{i,t}^2 = \sigma_{i,t-1}^2 + \frac{1}{n}\left(\frac{(n-1) * PE^2}{n} - \sigma_{i,t-1}^2\right) \qquad (5)$$

which allowed computation of how informative each new data point was, in terms of how much it could reduce variance about the local patch. As the above equation shows, the informativeness of each new data point decreased proportionally to *1/n* (see fig 1B). The model also tracked an estimate of the global mean, in terms of the history of visited patches. Each final mean estimate, μ_i, for the distribution within a patch served as an additional data point for inferring the grand mean M , in the same way that each within-patch data point served to estimate μ_i.

Crucially, the model assumed that upon first entering a new patch, the initial prediction regarding the distribution of that patch (essentially, the prior, before any data points from that patch were observed) was set to the current estimate of the global mean *M*. This provided a way to quantify the value of information in each patch, in terms of expected reward, as the estimated improvement in initial predictions on future patches. That is, the better the estimate of the global mean, the better the agent could do, on average, when entering a new patch. This is because the mean for each patch was drawn from a distribution centered on the global mean, and thus the optimal initial guess (prior) for a given patch was the global mean. Thus, at each time step *t*, the value of acquiring one extra data point in the current patch *i* could be estimated in terms of how much it improved future predictions (i.e. how much closer it moved them to *M*), relative to how much sampling a new patch would improve future predictions. Accordingly, we approximated the information value of staying in a patch with

$$V_{stay} = \frac{1}{(N-1)(n-1)} PE_{i,t} \qquad (6)$$

while the value of leaving the patch was

$$V_{switch} = \frac{1}{(N-1)} PE_{i+1,t} \qquad (7)$$

where *N* was the current number of patches exploited so far, *n* the current data points observed in the current patch, $PE_{i,t}$ the next estimated prediction error within the current patch, and $PE_{i+1,t}$ the next estimated prediction error assuming that the agent explored a new patch. This relative value between staying (exploiting) and switching (exploring) depended therefore on the current position within the game (*n*), the current position within the patch (*N*), and the two variance estimates for the patch mean (σ^2) and the global mean (S^2), as those variance estimates were used to compute the two prediction errors of interest in the above equation. Given a fixed number of available timesteps, the assumption that the agent could not return to a patch once it switched away, and values for the rewards *ρ*, *R* and the termination probability λ, we used dynamic programming to compute the value of staying or going at every time step (combining the immediate rewards for estimation within a patch with the approximate future value of learning the global mean, $\underline{V}$), as a function of how well both the local patch and the global mean were known.

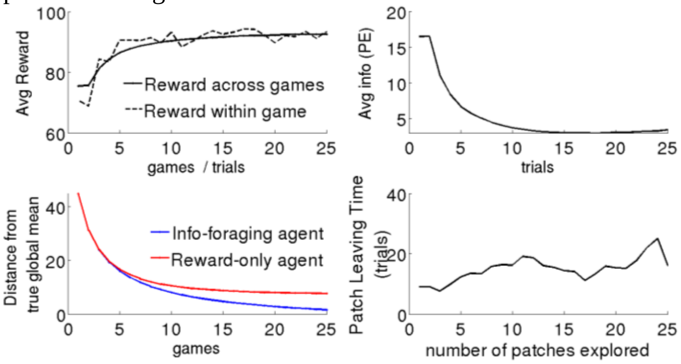

Figure 1: Model results. A. Reward increases with time spent in a game (dotted line), and average reward in a game increases as more patches are explored (solid black line). B. The amount of information obtained from each new sample decreases with trials spent in a game (i.e. patch). C. Information-foraging agent that leaves patches early (blue) learns the global structure parameters earlier than an agent that exploits a patch for all its reward (red). D. Model predicts later patch leaving times as a function of how well the world is known (i.e. how many patches have already been explored).

Results

Figure 1 shows model results (5000 simulations of twenty-five games each, with a maximum of twenty-five trials per game). As per the task structure, reward increases with time in game (fig. 1A, dotted line); furthermore the reward increases across games as global structure is learned (solid line). The change in prediction error (PE) as well as the reduction in uncertainty from each new data point decrease with time spent in game (fig 1B). Compared to an agent that exhausts all available trials in a patch, our "information-foraging" agent that leaves a patch depending on the relative informativeness of an extra data point within the patch versus a data point in a new patch showed faster learning of the global mean (i.e., learned it both in a shorter number of trials, and approached it faster across games, fig. 1C). Under certain model parameters, it also earned more average reward when compared to a model that only takes into account local reward (i.e. current prediction error). Looking at model predictions for the optimal time to switch away from a patch, as a function of both mean estimated variance of the global mean and, and as a function of time in game (though the two measures are somewhat correlated), the model predicted longer dwell times later in the game, when the uncertainty about the global mean had been significantly reduced (figure 1D).

Experiment 1

The following experiment was designed to test model predictions from our information-foraging model. The main prediction from the model, given the task structure presented to the subjects, was that they would quit high-reward, low-information patches early when the value of gaining information from new patches was higher (i.e., near the beginning of the task, when they had not learned much about the global environment), but spend longer and longer in patches as the usefulness of new information decreased.

Methods

Participants Twenty-five participants were recruited from the Princeton community. They gave informed consent, and were compensated for their time at a rate of $12/hour, plus a bonus of up to $5 for performance.

Task Participants played a game in which they controlled a virtual archer that made his way through enemy territory toward a castle (fig 2A, below). The goal was to defeat an "evil overlord," and the ability to do so could be enhanced by facing waves of "minions" trained by the overlord, and learning about their behavior as an indicator of his. Minions appeared sequentially on the screen, the archer had to fire an arrow to hit the minion, and doing so earned one point. A hit and miss counter was available on the bottom left of the screen, indicating to participants how well they were doing.

Participants were informed that the archer had to confront seven waves of minions before facing the overlord. Each wave consisted of a maximum of thirty-five minions, appearing one by one, from the right of the screen (fig 1A), at a variable location. Participants could adjust the archer's firing position on each trial, to best anticipate where the next minion would appear. At the end of seven waves, the archer confronted the overlord, and had only one shot to either defeat it (i.e., aim the arrow accurately enough to hit it), or be defeated by it (i.e. miss). A reward of 30 points was available for defeating the overlord.

Participants were informed that each wave of minions was trained by a different commander, and that all of the commanders had been trained by the overlord. Commanders shared, but did not perfectly imitate the overlord's preference for location of appearance. Similarly. minions shared, but did not perfectly imitate their commander's location of appearance. These instructions reflected the generative properties of the environment to be learned: the location of appearance of each minion within a wave was drawn from a distribution with a fixed mean and variance, and the means for each wave were drawn from a distribution with the same variance and a mean equal to the preferred location of the overlord.

Before encountering each minion in a wave, participants had two options: They could choose to stay and confront that minion, or choose to "run away" by pressing the large "RUN" button at the top left of the screen. If they chose to run, that wave of minions would end, a screen appeared announcing a new wave (with a new distribution of locations, and they would then have the same two options for each minion in the new wave. Given this task structure, sampling within a wave could lead to progressive improvement in performance (and reward) for a given wave. However, learning the preference of the overlord (associated with a much larger reward) required an appropriate balance of sampling within and across waves, as the informativeness of each data point decreased (see fig. 1B) within a wave, and other waves provided additional information.

Importantly, participants were told that they had only one hundred and fifty arrows to use on the minions – this operationalized the finite number of steps in our model – so they would have to decide how to use those arrows in a way that would give them the best shot of defeating the overlord. The model predicted that they should use fewer arrows (confront fewer minions) in earlier waves, switching waves (i.e., exploring) to maximize information about average locations of the waves (as a predictor of the overlord). At the same time, the model predicted that, as information accrued, and future opportunities to do so diminished (i.e., task horizon shortened), they should use arrows more liberally to earn points within each wave (i.e., exploit).

Results Twenty-five Participants learned the task, as evidenced both by their increasing accuracy in targeting the minions, within a wave (figure 2B) and by the increasingly accurate first location estimate – i.e., change of hitting the first minion – in later waves compared earlier waves (significant linear trend, $F(1,6) = 9.42$, $p = 0.02$, figure 2C).

No participants attempted to defeat all minions in a wave. However, participants' average number of minions attempted within a wave increased in later waves compared to earlier waves (figure 2C). This equates to earlier quitting times earlier in the game. Average likelihood to "run" (i.e. quit the current wave and move on to the next one) increased, for all participants, as a function of the number of minions they had confronted in the current wave (i.e the number of time steps they had spent there, fig. 2D).

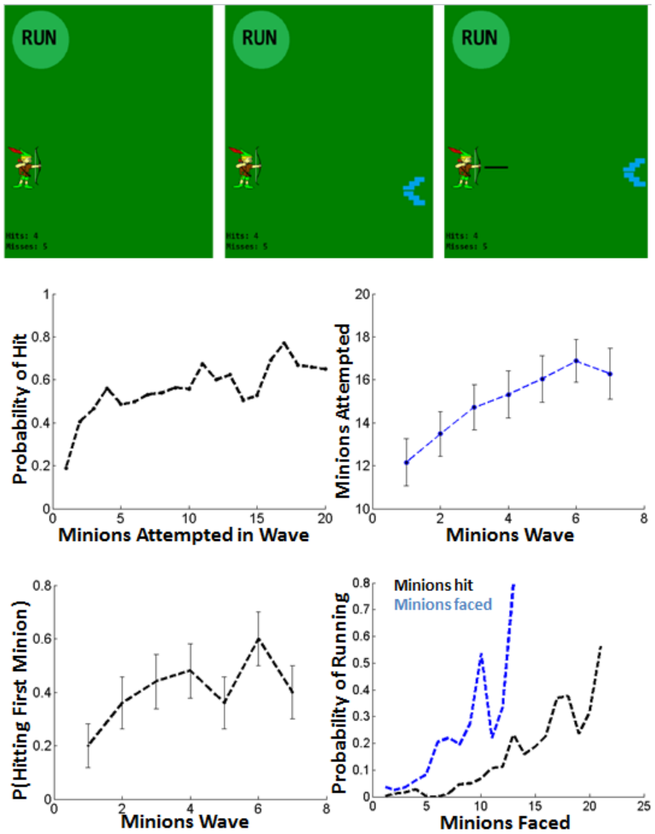

Figure 2: A. The archer task: participants adjust the position of the archer using the mouse, then release the arrow. The minion then comes from the right of the screen. B. Participants' chances of hitting a minion increase with time spent in a minion wave, as they learn to predict the locations of the minions better. C. Leaving times (i.e. the number of minions attempted) increase in later waves compared to earlier waves. D. The chance of hitting the first minion in a wave (i.e. the first sample of that particular wave's location) increases across the game. E. Participants are more likely to run after facing more minions (black line) and after hitting more minions (blue line).

Likelihood to run also increased with the number of minions actually defeated in the current wave (fig. 2E, blue line), which is well-correlated with how well the participants had learned that particular wave; comparatively, the likelihood to run was slower to increase as a function of faced minions (black line). Finally, in line with optimal performance, participants' accuracy in later waves improved on the first trial of a wave (before they got any data from the current minion wave), indicating that they generalized the knowledge about the structure of previous games to make better predictions in the current one (Fig. 2B).

Discussion

Building on ideas from optimal foraging theory and reinforcement learning, we proposed a normative model of adaptive exploration. The model is based on a variant of the Marginal Value Theorem (MVT; Charnov, 1976), in which explore-exploit decisions are driven by estimates of the relative information (rather than reward) associated with each option. The model balances the goals of learning about the properties of the local and global environments, and we showed that in doing so it is capable of optimizing overall reward rate. We tested, and found support for qualitative predictions of the model in an empirical study: human participants exhibited behavior consistent with information-seeking, and stay-leave (explore-exploit) decisions that were sensitive to estimates not only of current reward, but also current information, and that these were integrated into their decision to stay or leave.

The model predictions are consistent with theories of intrinsic motivation stating that the drive to explore arises from an innate need to interact efficiently with the environment (Deci & Ryan, 1985; White, 1959), as well as with the notion of "flow" and the optimal arousal theory of motivation, according to which organisms seek to balance an internal need for optimal levels of information and stimulation (Carrol, Zuckerman & Voegel, 1982; Csizentmihalyi, 2000). Furthermore, results here show that quitting a current high-reward but low-information patch can in fact still lead to higher overall reward than staying in the uninformative patch. This is consistent with model findings that average within-game reward increases across games, as the global structure is learned (Fig 1A). Similar findings have been presented previously in the machine learning literature, in studies showing that artificial agents capable of penalizing a too-well-known option's value can learn a complex grid environment faster and earn higher overall reward (Simsek & Barto, 2006). Ecological models of optimal foraging have also mentioned the "penalty of ignorance", i.e. the benefits that a forager could lose by not improving its information about the world over time (Olsson & Brown, 2006). However, to our knowledge, the model we present here is the first to cast exploration in terms that parallel the role of reward in optimal foraging theory. In this respect, the model provides a bridge between the closely-related literatures on foraging and exploration, and a path toward theoretical integration.

Along similar lines, our model provides a mechanistic, and potentially normative account of the phenomenology associated with boredom. The link between exploratory behavior and boredom has been suggested many times in both human and animal literature (Fowler 1959; Cohen, McClure & Yu, 2007; Meagher & Mason 2012; Kurzban, 2013). Our model formalizes this idea, suggesting that boredom might be considered as signaling the value of exploration; that is, that information provided by the current behavior is below what can be expected from alternatives, and therefore that a switch in behavior is warranted.

Consistent with this suggestion, we have found in a separate set of experiments that boredom is negatively correlated with the rate of information acquisition (learning) in the current task, is influenced by context, and that participants are willing to sacrifice reward in order to avoid boredom and increase the rate of information acquisition and learning. These observations are consistent with ones from the study reported here. At least initially, participants chose to switch away from a given wave of minions, even as their performance improved, reflecting a valuation of information and learning (and the diminution of "boredom") over immediate reward. This echoes a pervasive pattern of behavior in video games, and explains the need for "levels" to maintain gamers' interest — a phenomenon that is consistent with the current model. Interestingly, however, participants in our experiment chose to stay with a wave of minions *longer* as the task progressed; that is, they appear to have become *less* "bored" as overall time-on-task increased. This seemingly counterintuitive effect was predicted by the theory: as the task horizon shortened, the worth of information diminished relative to immediate reward, thus driving participants to persist (exploit) as the task neared an end.

It is important to note that, from the vantage of the model proposed, the value of exploration is dependent on the structure of the environment (e.g., the amount of time available, as well as the difficulty of the learning problem) and on how well the agent has learned the environment. Our findings apply to the pre-asymptotic phase(s) of learning, when gaining information from exploration can contribute to forming better representations of the environment and ultimately to better strategies for gaining reward. If the world were fully known, the same findings would not hold; however, given the complexity of the real world, it seems likely that organisms spend a considerable fraction of their time in pre-asymptotic phases of learning. That certainly appears to be the case for our collective understanding of how natural agents learn about their environment, and along similar lines, we hope that the work we describe here offers a useful new path for exploring this domain of understanding.

References

Behrens. T.E.J., Woolrich, M.W., Walton, M.E. * Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience* 10: 1214 – 1221

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem.*Theoretical population biology*, *9*(2), 129-136.

Cohen, J. D., McClure, S. M. & Yu, A. J. (2007).Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society: Biological Sciences. 362*, 933–942

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. Nature, 441, 876 – 879

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., & Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *The Journal of Neuroscience*, *33*(13), 5797-5805.

Eastwood, J. D., Frischen, A., Fenske, M. J., & Smilek, D. (2012). The Unengaged mind defining boredom in terms of attention. *Perspectives on Psychological Science*, *7*(5), 482-495.

Fowler, H. (1967). Satiation and Curiosity: Constructs for a Drive and Incentive-Motivational Theory of Exploration. *Psychology of learning and motivation*, *1*, 157-227.

Gallistel, C.R. (1990). *The Organization of Learning*. Cambridge, MA: Bradford Books/MIT Press.

Goldstone, R. L., & Ashpole, B. C. (2004). Human foraging behavior in a virtual environment. *Psychonomic bulletin & review*, *11*(3), 508-514.

Hill, A. B., & Perkins, R. E. (1985). Towards a model of boredom. *British Journal of Psychology*, *76*(2), 235-240.

Kaelbling, L., Littman, M., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237–285.

Krebs, J. R., Kacelnik, A. & Taylor, P. (1978) Tests of optimal sampling by foraging great tits. *Nature* 275, 27–31Lai & Robbins 1985

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(06), 661-679.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience*, *30*(37), 12366-12378.

Olsson, O., & S Brown, J. (2006). The foraging benefits of information and the penalty of ignorance. *Oikos*, *112*(2), 260-273.

Payzan-LeNestour, E., & Bossaerts, P. (2012). Do not bet on the unknown versus try to find out more: Estimation uncertainty and "unexpected uncertainty" both modulate exploration. Frontiers in Neuroscience

Pearson, J. M., Hayden, B. Y., Raghavachari, S., & Platt, M. L. (2009). Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multioption choice task. *Current biology*, *19*(18), 1532-1537.

Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In *From animals to animats: proceedings of the first international conference on simulation of adaptive behavior (SAB90)*.

Şimşek, O., & Barto, A. (2006). An intrinsic reward mechanism for efficient exploration. *Proceedings of the 23rd international conference*, 833-840.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.