Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration

Jonathan D. Cohen1,2,* , Samuel M. McClure1 and Angela J. Yu1

1Department of Psychology and Center for the Study of Brain, Mind and Behaviour, Princeton University, Princeton, NJ 08540, USA
2Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15213, USA

Many large and small decisions we make in our daily lives—which ice cream to choose, what research projects to pursue, which partner to marry—require an exploration of alternatives before committing to and exploiting the benefits of a particular choice. Furthermore, many decisions require re-evaluation, and further exploration of alternatives, in the face of changing needs or circumstances. That is, often our decisions depend on a higher level choice: whether to exploit well known but possibly suboptimal alternatives or to explore risky but potentially more profitable ones. How adaptive agents choose between exploitation and exploration remains an important and open question that has received relatively limited attention in the behavioural and brain sciences. The choice could depend on a number of factors, including the familiarity of the environment, how quickly the environment is likely to change and the relative value of exploiting known sources of reward versus the cost of reducing uncertainty through exploration. There is no known generally optimal solution to the exploration versus exploitation problem, and a solution to the general case may indeed not be possible. However, there have been formal analyses of the optimal policy under constrained circumstances. There have also been specific suggestions of how humans and animals may respond to this problem under particular experimental conditions as well as proposals about the brain mechanisms involved. Here, we provide a brief review of this work, discuss how exploration and exploitation may be mediated in the brain and highlight some promising future directions for research.

Keywords: exploration; uncertainty; learning; neurotransmitters; prefrontal cortex; decision making

1. INTRODUCTION

Should I stay or should I go now?
If I go there will be trouble
And if I stay it may be double
So come on and let me know
Should I stay or should I go?
(The Clash)

Every researcher has personal experience with the exploration–exploitation dilemma. At some point in the conduct of a study, when the data are still inconclusive, it may become necessary to decide how to proceed. On the one hand, there is the option to continue with the experiment, in the hope that with more effort and data, the results will look more promising. Alternatively, the experiment can be scrapped in favour of a modified experimental design, a new approach to the problem, or an entirely new research topic. That is, the experimenter faces a trade-off between the value of exploitation versus exploration. This example highlights the importance of this problem in decision making, one that has typically been ignored in psychological research on cognitive control and executive function.

The need to balance exploitation with exploration is confronted at all levels of behaviour and time-scales of decision making from deciding what to do next in the day to planning a career path. It is confronted by individuals in love (as captured by the lyrics above) and by entire armies at war (should a campaign focus intensively on one battle or seek to identify new opportunities to surmount the enemy). Nor is it limited to human behaviour. It is confronted by fungi deciding whether to concentrate growth at a local site or send out hyphae to sample more distant resources (Watkinson et al. 2005); by ant colonies exploring options for a new nest before settling on and exploiting a particular site (Pratt & Sumpter 2006); by engineers generating algorithms to deploy a fleet of automata to map the expanses of a new environment (Leonard et al. in press) and by machine learning theorists—who coined the phrase ‘exploration versus exploitation’—in their efforts to improve the ability of reinforcement learning (RL) algorithms to function adaptively in changing environments (e.g. Kaelbling et al. 1996).

In general, how agents should and do respond to the trade-off between exploration and exploitation is poorly understood. In part, this reflects the difficulty of the problem: there is no known optimal policy for trading off exploration and exploitation in general, even when the objectives are well specified. Gittins & Jones (1974) and Gittins (1979) presented a strategy and proved its
optimality for a limited class of problems in which the
decisions are made from a finite number of stationary
bandit processes (e.g. options for which the reward is
delivered with unknown but fixed probabilities), and
when the agent discount their value exponentially over
time. Gittins proved that if being optimal consists of
maximizing the cumulative reward over an infinite
horizon when the value of each reward is discounted
exponentially as a function of when it is acquired, then
the optimal policy is to calculate the expected total
future rewards associated with each option at a
particular time—a value known as the Gittins index—
and to select that bandit with the greatest Gittins index
(Gittins & Jones 1974; Gittins 1979). The significance
of Gittins’ contribution is that it reduced the decision
problem to computing and comparing these scalar
indices. In practice, computing the Gittins index is
not tractable for many problems for which it is known
to be optimal. However, for some limited problems,
explicit solutions have been found. For instance, the
Gittins index has been computed for certain two-
armed bandit problems (in which the agent chooses
between two options with independent probabilities of
generating a reward), and compared to the foraging
behaviour of birds under comparable circumstances;
the birds were found to behave approximately optimally
(Krebs et al. 1978).

While the Gittins index lends formal rigour to the
problem of exploration versus exploitation, proof of
its optimality requires strong assumptions about the
environment and the agent. The properties of the
individual bandits must be frozen unless acted upon
(i.e. the pay-off structure of the environment must be
stationary), all options must be available at all
decision points (i.e. there cannot be any ‘side
paths’) and agents must discount the value of rewards
exponentially into the future (Gittins 1979; Berry &
Fristedt 1985; Banks & Sundaram 1994). Real-world
problems typically violate one or more of these
assumptions.

Perhaps, the most important exception to Gittins’
assumptions is that real-world environments are
typically non-stationary; i.e. they change with time.
To understand how organisms manage the balance
between exploration and exploitation in non-stationary
environments, investigators have begun to study how
organisms adapt their behaviour in response to the
experimentally induced changes in reward contingencies.
Several studies have now shown that both humans
and other animals dynamically update their estimates
of rewards associated with specific courses of action,
and abandon actions that are deemed to be diminishing
in value in search of others that may be more rewarding
(e.g. Sugrue et al. 2004; Daw et al. 2006; Gilzenrat &
Cohen in preparation). At the same time, there is
also longstanding evidence that humans sometimes
exhibit an opposing tendency. When reward diminishes
(e.g. following an error in performance), subjects often
try harder at what they have been doing rather than less
(e.g. Rabbitt 1966; Laming 1979; Gratton et al. 1992).
The balance between exploration and exploitation also
seems to be sensitive to time horizons. Humans show
a greater tendency to explore when there is more
time left in a task, presumably because this allows
them sufficient time later to enjoy the fruits of
those explorations (Carstensen et al. 1999). A full
account of how people regulate the balance between
exploration and exploitation must account for these
diverse, and in some cases seemingly discrepant,
patterns of behaviour.

Recent findings are also beginning to shed light on
the neural mechanisms that underlie exploratory and
exploitative behaviours. These findings consistently
implicate the involvement of neuromodulatory systems
thought to be involved in assessing reward and
uncertainty. The midbrain dopamine system has been
implicated in the signalling of reward prediction errors
critical for learning the value of specific actions
(Montague et al. 1996; Schultz et al. 1997) and for
decision-making based on those values (McClure et al.
2003). The locus coeruleus (LC) noradrenergic system
has been proposed to govern the balance between
exploration and exploitation in response to reward
history (Aston-Jones & Cohen 2005). And the basal
forebrain cholinergic system together with the adre-
nergic system have been proposed to monitor uncer-
tainty, signalling both expected and unexpected forms,
respectively, which in turn might be used to promote
exploitation or exploration (Yu & Dayan 2005).

Regulating the balance between exploitation and
exploration is a fundamental need for adaptive
behaviour in a complex and changing world. In the
rest of this article, we consider the progress outlined
above that has been made in understanding this
problem in formal terms and in identifying the
mechanisms that have evolved in natural organisms
for meeting this challenge. While there has been recent
progress in identifying relevant empirical phenomena
and candidate neural mechanisms, such work is still in
the earliest stages. Accordingly, the connection
between theory and data remains largely speculative.
Our primary purpose here is to call attention to the
problem and point to relevant lines of research that
show promise in addressing it.

2. OPTIMAL PERFORMANCE IN STATIONARY
ENVIRONMENTS: THE GITTINS INDEX
In a landmark paper, Gittins & Jones (1974) developed
a straightforward means for calculating the optimal
strategy for decision making in multi-armed bandit
problems. Bandit problems are well suited for studying
the tension between exploitation and exploitation since
they offer a direct trade-off between exploiting a known
source of reward (continuing to play one arm of the
bandit) and exploring the environment (trying other
arms) to acquire information about other sources of
reward (Kaelbling 1996).

For an $n$-armed bandit problem, an agent is required
to choose between $n$ options, each of which delivers
reward with a probability $p_i$. The probability of
obtaining reward from a bandit, $p_i$, may change through
time but only when a choice is made for that bandit. The
goal for the agent is to maximize expected rewards, $V$, where
rewards earned in the future are discounted by an
exponential discount factor $\delta \in (0, 1)$.
Gittins & Jones proved that optimal performance can be obtained by tracking a single index $v_i$ of the form

$$v_i = \sup_{T > 0} \left( \frac{\sum_{t=0}^{T} \delta^t R_i(t)}{\sum_{t=0}^{T} \delta^t} \right),$$

(2.1)

for each of the bandits, which is a normalized sum of future rewards discounted by the delay until they are accrued. The sum is taken until a time $T$, which is defined as the stopping time, or the point at which selecting from bandit $i$ will be terminated. Gittins & Jones proved that optimal behaviour is assured as long as that action is always taken which has the greatest index value. Critically, the Gittins index for any given bandit is independent of the expected outcomes of all other bandits. This implies that once the bandit with greatest index is known, behaviour should continue on this bandit until its index value falls below its original value. This is true because the index values for all other bandits do not change as long as these bandits are not selected. Computationally, calculating the Gittins index (equation (2.1)) is demanding and may not reasonably be expected to be calculated in the brain.

The Gittins index provides a normative account of how agents should act when faced with a particular form of the exploration–exploitation dilemma. Krebs et al. (1978) tested whether the foraging behaviour of birds is optimal when confronting a two-armed bandit problem similar to that solved by the Gittins index. In the experiment, the birds were presented with two feeding posts that gave food reward with fixed probability. The problem was a simplification of the general problem solved by the Gittins index, since the probability of obtaining reward from a feeding post was not allowed to change when selected and since the experiment was of finite length. The investigators found that the time at which birds stopped exploring (operationalized as the point at which they stayed at one feeding post) closely approximated that predicted by the optimal solution. Despite their findings, Krebs et al. (1978) recognized that it was highly unlikely that their birds were carrying out the complex calculations required by the Gittins index. Rather, they suggested that the birds were using simple behavioural heuristics that produces exploration times that qualitatively approximate the optimal solution. However, there are more fundamental problems with the Gittins index, beyond complexity of calculation.

As noted earlier, Gittins’ proof requires that rewards should be discounted exponentially for delay (Berry & Frisestd 1985) whereas it is generally accepted that most animals (including humans) show hyperbolic discounting (e.g. Ainslie 1975). Additionally, if there is a cost associated with switching from one behaviour to another, then not only is the Gittins index no longer optimal, but also there is no optimal index that may be calculated independently for each bandit (Banks & Sundaram 1994). It is well recognized that, under many conditions, humans exhibit costs when switching from one task to another (e.g. Allport et al. 1994; Rogers & Monsell 1995). Most importantly, the Gittins index assumes that, although the pay-offs for each bandit are probabilistic and each must be sampled sufficiently to determine its expected value, the actual expected value of each remains fixed except when acted upon. That is, if nothing is done to a bandit, then its true value remains stable across time. However, both the needs of most organisms and the environments in which they live are not stable in this way. Things change over time, even when they are not acted upon, and often in unpredictable ways. To date, no universally optimal algorithm has been described that prescribes how to trade-off between exploration and exploitation in non-stationary environments, and it is not clear that doing so is possible. Thus, understanding how animals respond to this problem must also be guided by empirical investigation, both of behaviour and underlying neural mechanisms.

3. MODELLING EXPLOITATION VERSUS EXPLORATION IN NON-STATIONARY ENVIRONMENTS

Daw et al. (2006) recently addressed this problem in a study that used a variant of the $n$-armed bandit problem in which the pay-offs of each bandit changed slowly over time (figure 1). In this setting, therefore, the cost of persisting with one behaviour (i.e. playing only one bandit) was not only the opportunity cost of failing to learn more about the value of the others, but also the possibility that what has already been learned about them will fall out of date. Daw et al. (2006) proposed three possible models for how subjects might guide their choices in this situation.

The first model used a simple decision rule, in which the subject maintains a record of the expected value for each option, based on past experience, and usually chooses the option with the greatest value (exploitation) though sometimes, with a fixed probability, picks randomly among the other alternatives (exploration). This is often referred to as the ‘epsilon-greedy’ algorithm (Sutton & Barto 1998). According to a second model, options are chosen by probability matching, i.e. with a probability weighted by their estimated values. This is often referred to as the ‘soft max’ decision rule (e.g. Thrun 1992), as it favours choosing the option with the maximum value (this option will have the highest probability), though this tendency is ‘softened’ by both the value of the competing options as well as randomness (noise).
added to the decision rule. Thus, in this model, the balance between exploitation and exploration is governed by both the relative value of the alternatives as well as a parameter (referred to as gain or, inversely, temperature) that determines how tightly decisions are constrained by the contrast of value among the alternatives: with higher gain, decisions are determined more by relative value (exploitation); with lower gain, decisions are more evenly distributed at random (exploration).

Finally, they entertained a third model, according to which choices are made using the soft max decision rule, but with a critical added factor: options that have not been selected receive an ‘uncertainty bonus’ that augments their probability of being chosen (i.e., promotes exploration). This captures the opportunity cost that is formalized by the Gittins index for stationary environments, and that is particularly important in non-stationary environments: the more time allocated to one option the less one knows about the others, which may be (or have become) more valuable.

Daw et al. (2006) compared the behaviour of subjects playing their n-arm bandit task to predictions from each of the three models. The model that provided the best fit was the soft max decision rule. Importantly, although subjects did periodically explore options other than the one currently deemed to be most valuable, they did not find evidence that this was driven by an uncertainty bonus (i.e., growing uncertainty about the competing alternatives). However, there are several caveats that must be kept in mind. First, it is possible that the specifics of the environment did not adequately favour the use of an uncertainty bonus. For example, the pay-offs of each bandit changed continuously and relatively slowly over time in their experiment. In the real world—to which real-world organisms are presumably adapted—the dynamics of environmental change may be very different, and therefore call for a different policy of exploration (and computation of uncertainty bonus) than was assumed by Daw et al. (2006). Another important factor may be social context—people may be enticed to explore the environment when they have information about the behaviour of others, and they may also place a greater premium on exploration when they face competition from others for resources.

These are questions that beg more detailed formal analysis. Nevertheless, to our knowledge, the Daw et al. (2006) study was the first to address formally the question of how subjects weigh exploration against exploitation in a non-stationary, but experimentally controlled environment. It also produced some interesting neurobiological findings. Their subjects performed the n-armed bandit task while being scanned using functional magnetic resonance imaging (fMRI). Among the observations reported was task-related activity in two sets of regions of prefrontal cortex (PFC). One set of regions was in ventromedial PFC and was associated with both the magnitude of reward associated with a choice, and that predicted by their computational model of the task (using the soft max decision rule). This area has been consistently associated with the encoding of reward value across a variety of task domains (O’Doherty et al. 2001; Knutson et al. 2003; McClure et al. 2004; Padoa-Schioppa & Assad 2006). A second set of areas observed bilaterally in frontopolar PFC was significantly more active when subjects chose to explore (i.e., chose an option other than the one estimated by their model to be the most rewarding) rather than exploit. This finding is consistent with the hypothesis that more anterior and dorsal regions of PFC are responsible for top-down control, biasing processes responsible for behaviour in favour of higher level goals, especially when these must compete with otherwise prepotent behaviours (e.g., Miller & Cohen 2001). Such top-down control may be important for exploration, insofar as this involves selecting an action that has been less recently associated with reward. That is, when a decision is made to pursue an exploratory behaviour, this may rely on support from higher level control processes. However, this begs the question of how the system decides when it is appropriate to explore. That is, what mechanisms are responsible for assessing the reliability and value of current rewards, and using this information to determine when to continue to pursue current sources of reward (exploit) or take a chance in pursuing new behaviours (explore). Several lines of investigation have begun to address this question.

4. Uncertainty and exploitation versus exploration
One line of work that has direct relevance addresses the question of how the brain encodes different forms of uncertainty. Yu & Dayan (2005) proposed that a critical function of two important neuromodulators—acetylcholine (ACh) and norepinephrine (NE)—may be to signal expected and unexpected sources of uncertainty. While the model they developed for this was not intended to address the trade-off between exploitation and exploration, the distinction between expected and unexpected uncertainty is likely to be an important factor in regulating this trade-off. For example, the detection of unexpected uncertainty can be an important signal of the need to promote exploration. To see this, consider the following scenario.

You are asked to observe a series of coin tosses, told that the coin is biased, and your job is to determine whether it is biased towards heads or tails. The first few coin tosses produce the following sequence: heads, heads, tails, heads, heads, heads, heads, heads, tails, tails, tails, tails... At some point, you will revise your determination and say that the coin is biased towards tails. Perhaps, the coin was surreptitiously switched (i.e., the world has changed) or your determination was wrong in the first place. In either case, having come to assume that the coin is biased towards heads, you have now been confronted with an unexpected form of uncertainty and must revise your model of the world accordingly, along with the choice of any actions that depend on it.

This problem is closely related to the example we gave at the beginning of this article (concerning the
collection of experimental data), and as we have noted elsewhere (Aston-Jones & Cohen 2005), the distinction
between expected and unexpected forms of uncertainty may be an important element in choosing between
exploitation versus exploration. As long as prediction
erors can be accounted for in terms of expected
uncertainty—that is the amount that we expect a given
outcome to vary—then all other things being equal
(e.g. ignoring potential non-stationarities in the
environment), we should persist in our current
behaviour (exploit). However, if errors in prediction
begin to exceed the degree expected—i.e. unexpected
uncertainty mounts—then we should revise our
strategy and consider alternatives (explore).

Yu & Dayan (2005) proposed that ACh levels are
used to signal expected uncertainty, and NE to signal
unexpected uncertainty. They describe a computationally tractable algorithm by which these may be estimated
that approximates the Bayesian optimal computation of
those estimates. Furthermore, they proposed how these
estimates, reflected by NE and ACh levels, could be
used to determine when to revise expectations

\[ \text{NE} > \frac{\text{ACh}}{0.5 + \text{ACh}}. \]  

(4.1)

They showed that this closely approximates the
Bayesian optimal solution to, and people’s behaviour
in, a variant of a commonly used selective attention task
(the ‘Posner paradigm’; Posner et al. 1980).

This work provides another instructive example of
the value in conducting a mathematical analysis of
optimal performance in a task, and using this to guide
the generation of hypotheses about the specific
mechanisms—in this case neural—that govern
behaviour in that task. Furthermore, it lends precision
to hypotheses about the function of neuromodulatory
systems. Despite their ubiquity in the brain, theories
about these systems have typically been vague,
proposing non-specific functions such as the mediation
of motivation and arousal. Yu and Dayan’s model assigns
precise functions to ACh and NE, specified in
mathematical form, that can be used to generate
specific testable predictions.

As suggested above, it is not hard to imagine how the
functions ascribed to ACh and NE in representing
estimates of expected and unexpected forms of
uncertainty might play an important role in regulating
the balance between exploitation and exploration. As
estimates of unexpected uncertainty rise, and NE
approaches the threshold defined by equation (4.1),
the system promotes a revision of current expectations.
This could be an important signal to search for a new
model of the environment and a corresponding
behavioural strategy—i.e. exploration. Sometimes,
however, unexpected events are followed by the
opposite tendency: an increase in commitment to the
current behavioural strategy. For example, following
errors in simple reaction time tasks people often
become more cautious and improve their performance
(i.e. become more accurate; Rabbitt 1966; Laming
1979). Similarly, following interference in selective
attention tasks, subjects typically increase the focus of
their attention and improve performance (Gratton et al.
1992), especially when such interference is relatively

The Yu & Dayan model also sensibly predicts that
performance should be best when expectations are most
accurate. However, when outcomes in a task become too
predictable, people often become bored and look for
other things to do (explore). Video game programmers
learned this lesson long ago, and routinely include
multiple levels in a game, so that when it becomes too
predictable, it is made more difficult in order to retain
players’ interest (i.e. keep them exploiting).

These observations suggest that additional
mechanisms may be involved in evaluating expectations
and in regulating the trade-off between exploration and
exploitation. Another closely related line of investigation
has sought to address some of these observations.
It too has suggested an important role for NE, building
on detailed physiological observations about the
dynamics of NE release, and proposing how this may
relate to assessments of reward as well as uncertainty.

5. UTILIT Y AND EXPLOITATION
VERSUS EXPLOR A TION

Virtually all of the NE released in the neocortex originates
from a small brainstem nucleus called the LC. Aston-
Jones et al. (1994, 1997) have observed that in the awake
behaving monkey the LC shifts between two operating
modes that correspond closely with behavioural
performance in a simple target detection task. In the ‘phasic
mode,’ when the animal is performing well (no misses
and very few false alarms), the LC shows only moderate
levels of tonic discharge, but responds phasically with
a burst of activity to target stimuli (but not to distractors).
In the ‘tonic mode’, the baseline level of discharge is
higher, but there are diminished or absent phasic
responses to target stimuli. In this mode, reaction time
to targets is slower and the animal commits a greater
number of false alarms to distractors. These two modes
most probably represent a continuum of LC function,
consistent with the formal theories described below.
However, we will continue to refer to two modes for
expository purposes, because the distinction between
them (or the extremes of function they represent) has
been proposed to be an important factor in influencing
the balance between exploration and exploitation.

Usher et al. (1999) developed a biophysically
detailed model of the LC that accounted for the
physiological observations outlined above and
suggested that these may play a role in regulating
the balance between exploitation and exploration. They
proposed that the phasic mode favours exploitation by
releasing NE specifically when a task-relevant event
occurs, thereby facilitating processing of that event. In
contrast, in the tonic mode, sustained release of NE
indiscriminately facilitates processing of all events
irrespective of their relevance to the current task and
thereby favours exploration. Note that the latter aligns
well with the role of NE proposed by Yu & Dayan
(2005), favouring exploration, if it is assumed that NE
in their model corresponds to tonic release.

The Usher et al. (1999) model describes physiological
mechanisms by which the LC may contribute to
regulating the balance between exploitation and
exploration. However, it does not specify what drives the LC towards the phasic (exploitation) or tonic (exploration) modes. Recently, Aston-Jones & Cohen (2005) have proposed that this may be governed by ongoing assessments of utility carried out in ventral and medial frontal structures. As noted earlier, there is extensive evidence that ventral regions within PFC form part of a circuit responsible for encoding reward value (e.g. Knutson et al. 2003; O’Doherty et al. 2001; McClure et al. 2004; Padoa-Schioppa & Assad 2006). There is also now a substantial body of evidence that medial frontal structures, and in particular the anterior cingulate cortex (ACC), encode costs. Regions within the ACC have consistently been observed to respond to pain, negative feedback, errors in performance, conflicts in processing and even mental effort, all of which represent or are indicative of various forms of cost (e.g. Milner et al. 1997; Carter et al. 1998; Peyron et al. 2000; Botvinick et al. 2001; Holroyd & Coles 2002; Yeung et al. 2004). Furthermore, recent anatomic evidence indicates that these ventral and medial frontal structures provide dense projections to the LC (Rajkowski et al. 2000; Aston-Jones et al. 2002).

Based on these findings, Aston-Jones & Cohen (2005) have proposed that ongoing assessments of utility carried in frontal structures are used to govern the mode of LC and thereby regulate the balance between exploitation and exploration. Specifically, they propose that assessments of utility are carried out over both short (e.g. seconds) and long (e.g. minutes) time-scales and that this can reconcile the opposing tendencies (to ‘try harder’ versus ‘give up’) following periods of poor performance noted above. For example, consider the following two circumstances. In one, performance in a task has been good and there are still rewards to be accrued from the task, but there are occasional lapses in performance producing transient decreases in utility (e.g. on single trials). In this case, following such a lapse the agent should act to restore performance. That is, exploitation should be promoted when long-term utility has been high, but there has been a momentary decrease. In contrast, consider a second situation in which performance has been poor and utility has progressively declined. At some point, this should encourage disengagement from the current task and exploration of alternative behaviours. That is, how the system responds to a current decrease in utility should depend upon the context of longer term trends in utility, favouring exploitation if long-term utility has been high, and exploration if it has been low. A relatively simple equation can capture these relationships,

\[
\text{Engagement in current task} = [1 - \text{logistic}(\text{short-term utility})] \\
\times [\text{logistic}(\text{long-term utility})],
\]

where logistic refers to the sigmoid function \(1/(1 + e^{-u_{\text{unity}}})\).

Aston-Jones & Cohen (2005) proposed that high values of this equation favour the LC phasic mode (exploitation), whereas low values favour the tonic mode (exploration; figure 2). Usher et al. (1999) and Brown et al. (2005) both suggest the ways in which this can be accomplished through the regulation of simple physiological parameters (such as electronic coupling and/or baseline afferent drive) within the LC.

This model can also be related to the soft max mechanism that Daw et al. (2006) found best fits decision-making behaviour in their \(n\)-armed bandit task. The effect of the LC can be thought of as tuning the softmax function, sharpening it (phasic mode) and biasing decisions towards the most recently rewarded choices (i.e. exploitation) when long-term utility is high, and flattening the function (tonic mode) promoting a more uniform distribution of choices (exploration) when long-term utility is low. Whether such effects are observed in a suitably designed \(n\)-armed bandit decision-making task remains to be tested. However, recent findings from a simpler, two-armed decision-making task, that used pupillometry to index LC activity (Aston-Jones & Cohen 2005), have corroborated predictions of the model regarding the relationship of LC activity to decision-making performance (Gilzenrat & Cohen in preparation). This work has also recently been extended to explore the interaction between these mechanisms and those underlying RL.

6. REINFORCEMENT LEARNING AND EXPLOITATION VERSUS EXPLORATION

The trade-off between exploration and exploitation has long been recognized as a central issue in RL (Kaelbling 1996, 2003). The RL mechanisms act by strengthening associations (e.g. between a stimulus and an action) when these have been associated with a reward (e.g. Sutton & Barto 1998). There is now strong reason to believe that the dopaminergic (DA) system implements such a mechanism (Montague et al. 1996; see Montague et al. 2004 for a recent review). The RL mechanisms function well in stationary environments, in which progressive strengthening of associations makes them robust and efficient, allowing the agent to exploit the current environment. However, this also makes them resistant to change, which is problematic.
conditioning paradigm using a target detection task (Aston-Jones et al. 1997). Performance of the task following acquisition of the initial target was associated with the LC phasic mode. When the contingencies were reversed, LC tonic activity increased and phasic responses diminished. Then, as the new target was acquired, the LC returned to the phasic mode of responding. These findings provide growing support for the view that the LC noradrenergic system plays an important role in mediating the balance between exploitation and exploration. As the work of Aston-Jones & Cohen (2005) and Yu & Dayan (2005) suggests, ongoing assessments of both uncertainty and utility are likely to be important in regulating this balance.

7. OPEN QUESTIONS AND CHALLENGES

In this article, we hope to have drawn attention to the fact that managing the trade-off between exploitation and exploration is a fundamental challenge for the adaptive control of behaviour. While traditionally this has not occupied centre stage in research on executive function and cognitive control, we have reviewed several lines of work that have productively begun to address this issue. Nevertheless, many important questions remain.

First, it is should be noted that some of the work we have reviewed addresses the estimation of uncertainty (e.g. Yu & Dayan 2005), while other work focuses more on the computation of utility and action selection (e.g. Usher et al. 1999; Aston-Jones & Cohen 2005; Daw et al. 2006). All of these are likely to be critical elements in determining the trade-off between exploitation and exploration. However, the specific relationship between these remains to be examined directly. For example, it would be valuable to understand how the mechanisms proposed by Yu & Dayan (2005) to compute assessments of uncertainty (i.e. prediction errors) can be coupled to action selection, and how this relates to the algorithm described by equation (5.1)—proposed by Aston-Jones & Cohen (2005) to relate assessments of utility to LC function and decision-making performance.

It seems inescapable that, in addition to uncertainty and utility, social signals are a critical factor adjudicating the trade-off between exploitation and exploration. Observing others can provide critical counterfactual information about the reward value of behavioural strategies that one has not yet pursued oneself (Montague et al. in press). Competition within a social context may also help explain aspects of boredom—i.e. the perplexing tendency to explore alternatives to current behaviour when certainty of outcome (including reward) is at its highest. If it is assumed that more difficult tasks are both more remunerative and less competitive (because fewer agents possess the skills necessary to perform them), then performing a task below one’s skill level carries an opportunity cost. That is, it should be possible to find another task for which one is still adequately competent, but that is more difficult and less competitive, and therefore more remunerative. Thus, boredom may in part reflect an adaptive bias towards exploration when performance at ceiling suggests that a more remunerative task can be found (M. Todd 2006, personal communication).
An important super-ordinate question is whether the trade-off between exploitation and exploration should be considered a single problem addressed by a unitary set of mechanisms in the brain, or whether it represents a family of problems spanning different scales, that are addressed by different mechanisms. The time-scale of neuromodulatory function suggests that these mechanisms influence decisions that take place over seconds or minutes. However, faster processes (e.g. saccadic search mechanisms) and longer ones (planning a career) may involve very different mechanisms.

Finally, an equally pressing question is whether it is best to distinguish qualitatively between exploitation and exploration, or whether these represent the extremes of a continuum. For example, the models we have discussed have, for the most part, treated exploration as random search (e.g. increasing noise in an annealing procedure). However, search can often be structured by relatively sophisticated, domain-specific heuristics (for example, in problem solving tasks; Newell & Simon 1972). Such search processes may involve temporally extended, goal-directed behaviours that rely on mechanisms of cognitive control similar to those required for exploitation within the context of simpler tasks. Indeed, the findings of an association between PFC activity and exploration in the Daw et al. (2006) study may provide an example of this. These considerations help underscore the need for a precise formulation of the exploitation–exploration trade-off within specific task environments.

More generally, this issue brings into focus an important dimension for considering the trade-off between exploitation and exploration: the extent to which the environment to be explored is well-structured (whether static or changing in predictable ways) versus unknown and unpredictable. To the extent that it is structured, then it should be possible to explore it in a systematic fashion (we might refer to this as ‘controlled exploration’). That is, at least from the theorist’s perspective, it should be possible to identify an optimal strategy for exploration that takes account of knowledge about the various behavioural alternatives (the Gittins index represents a special case of this). Under such conditions, the decision of whether to exploit or explore should weigh both the value of current pursuits as well as informed expectations about the alternatives, and exploration should be deterministic. Indeed, to the extent that an optimal strategy can be found, this might be thought of simply as higher level exploitation. Of course, for realistically complex environments, theoretically optimal strategies are likely to be computationally intractable, at least for biological mechanisms (this is so for the Gittins index, even given its simplifying assumptions). Thus approximations, including stochastic ones (such as some of the mechanisms reviewed in this article) may be more biologically realistic.

At the other end of this dimension are unknown and unpredictable environments. Under such conditions, the decision of whether to exploit or explore may focus more profitably on assessments of performance in the current task rather than on expectations about alternatives. Similarly, strategies for exploration will necessarily rely on cruder assumptions about behavioural alternatives and search among them will be less structured and presumably more stochastic. Consideration of these factors may be useful in guiding the next generation of hypotheses about the mechanisms governing exploitation and exploration in biological organisms.

8. SUMMARY AND CONCLUSIONS
This article began by reviewing efforts to formalize the optimal solution to the trade-off between exploitation and exploration. The Gittins index provides such a solution, but applies to restricted circumstances (e.g. only stationary environments). As yet, no general solution has been found for non-stationary environments and, depending upon the breadth and characteristics of the environment to be considered, this may not be possible. Nonetheless, empirical studies of both behaviour and neural mechanisms have begun to reveal mechanisms that animals may use to adapt to changes in the environment, by regulating the balance between exploitation and exploration. These studies appear to be converging on the view that neuromodulatory systems—in particular, ACh and NE, interacting with DA-mediated RL mechanisms—may play a critical role in regulating this balance within certain domains of behaviour. These systems appear to be responsive to both estimates of uncertainty and utility. However, social signals are also likely to be an important source of information. More generally, the trade-off between exploitation and exploration represents a challenge to behaviour at all levels and over multiple time-scales. It is not yet clear whether neuromodulatory mechanisms serve the same function at all of these levels and timescales, or whether this relies on other mechanisms that remain to be discovered. Given these considerations, it seems probable that further research will require a mixed (though not yet fully informed) strategy of continuing to exploit promising lines of recent work, while considering new ones to explore.

This work was supported by NIH grants P50 MH062196 (J.D.C.), F32 MH072141 (S.M.M.) and the NIMH Quantitative Neuroscience Training Grant (MH65214). We thank Gary Aston-Jones, Eric Brown, Mark Gilzenrat, Phil Holmes and Leigh Nyström for their close collaboration in the development of many of the ideas presented in this article. We would also like to thank Nathaniel Daw, Yael Niv and Greg Stephens for their valuable discussions related to this work, as well as Peter Dayan and an anonymous reviewer for their useful suggestions regarding this article. Finally, we would like to offer a profound and heartfelt thanks to Tim Shallice, not only for his comments on this article, but more importantly for the decades of inspiring work and visionary leadership that he has provided to our field.

REFERENCES


