Opinion TRENDS in Cognitive Sciences Vol.11 No.12

Full text provided by www.sciencedirect.com
ScienceDirect

Neuroeconomics

Efficient statistics, common currencies and the problem of reward-harvesting

P. Read Montague and Brooks King-Casas

Department of Neuroscience and Computational Psychiatry Unit, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA

The mammalian brain is equipped with reward-harvesting mechanisms that efficiently categorize and value the behavioral choices that lead to rewards necessary for survival. In this context, 'efficiency' embodies the idea of achieving maximum returns for minimal energetic investments and places a premium on how an animal represents its available options. But the capacity to efficiently represent choices is a profoundly difficult problem because representations for behavioral choice depend intimately on the statistics of information arriving not just from the sensory world and but also from within the creature itself. Any complete account of decision-making in mammals must efficiently connect the internal needs to the perceptual apparatus available to a creature moment-to-moment.

Introduction

> *"Instead of thinking of neural representations as transformations of stimulus energies, we should regard them as approximate estimates of the probable truths of hypotheses about the current environment"* (Horace Barlow, 2001)

Mobile organisms run on 'batteries', so they are forced to be rapid-fire economic decision-makers that know how to value their past, present and future [1,2]. These kinds of introductory phrases are easy to make because they are sufficiently vague and it is generally accepted that any decision-maker must have some way to differentially value its world. Consequently, most summaries of the problems taken on by neuroeconomics start with such slogans and progress quickly to whole-organism decision-making and its relation to some neural probe of choice – usually functional magnetic resonance imaging or single-unit electrophysiology. This short review will end up being guilty of that same sequence of events, but it is worth considering briefly the foundations and implicit assumptions on which such approaches are founded, because these foundations also point the way to the future of neuroeconomics. The key idea that connects these introductory slogans to underlying neural components is that of 'batteries'.

The natural stimulus statistics hypothesis

For real-world organisms, batteries (limited lifetime energy supplies) have forced innumerable energetic constraints on the way that biological brains operate, from the level of single molecules all the way up to algorithms for behavioral control [3–5]. These forced 'efficiencies' have long invited information-theoretical analysis of neural processing, an approach that is now approaching the end of its sixth decade.

As early as 1952, Donald MacKay and Warren McCulloch [6] employed 'considerations of efficiency' to determine the performance limits for information transmission at synaptic junctions. Their study displayed a remarkably modern information-theoretical outlook. Visual neuroscientist Horace Barlow upgraded this same information-theoretical perspective to a hypothesis about the efficiencies of neural function that one should expect on the basis of 'demands' made by the sensory world. In 1961, he proposed what could now be called the 'natural stimulus statistics hypothesis', which posits that the visual system possesses its specific features because it is adapted to the demands imposed by the statistics of real-world visual scenes [7]. This claim might seem obvious at first; however, it was a singular step in our understanding of the likely nature of neural representations underlying perception – in this case visual perception. The import of Barlow's clearly stated idea was that it provided a way to generate 'guided hypotheses' about what particular visual functions were 'for'; this is meant in the same spirit as the idea that the heart is 'for' pumping blood [8].

Barlow pointed out that visual scenes contain loads of redundant information; they possess lots of local correlations in time and space. An efficient nervous system should not have to process all these redundancies, but should have evolved ways to reduce them. Redundancy-reduction mechanisms could take many specific forms; however, their action will tend to generate statistically independent (de-correlated) representations of the visual world (Figure 1a). More than twenty years later, Barlow and Foldiack [9] used this same perspective to hypothesize how moment-to-moment de-correlation through time and tissue space could be implemented by real neurons and how such dynamic de-correlation mechanisms would steer visual responses toward statistical independence. Presently, many decades after Barlow's initial proposal, this general outlook has blossomed into a principled way to

Corresponding author: Montague, P.R. (read@bcm.tmc.edu).

 1364-6613/$ – see front matter © 2007 Elsevier Ltd. All rights reserved. doi:10.1016/j.tics.2007.10.002

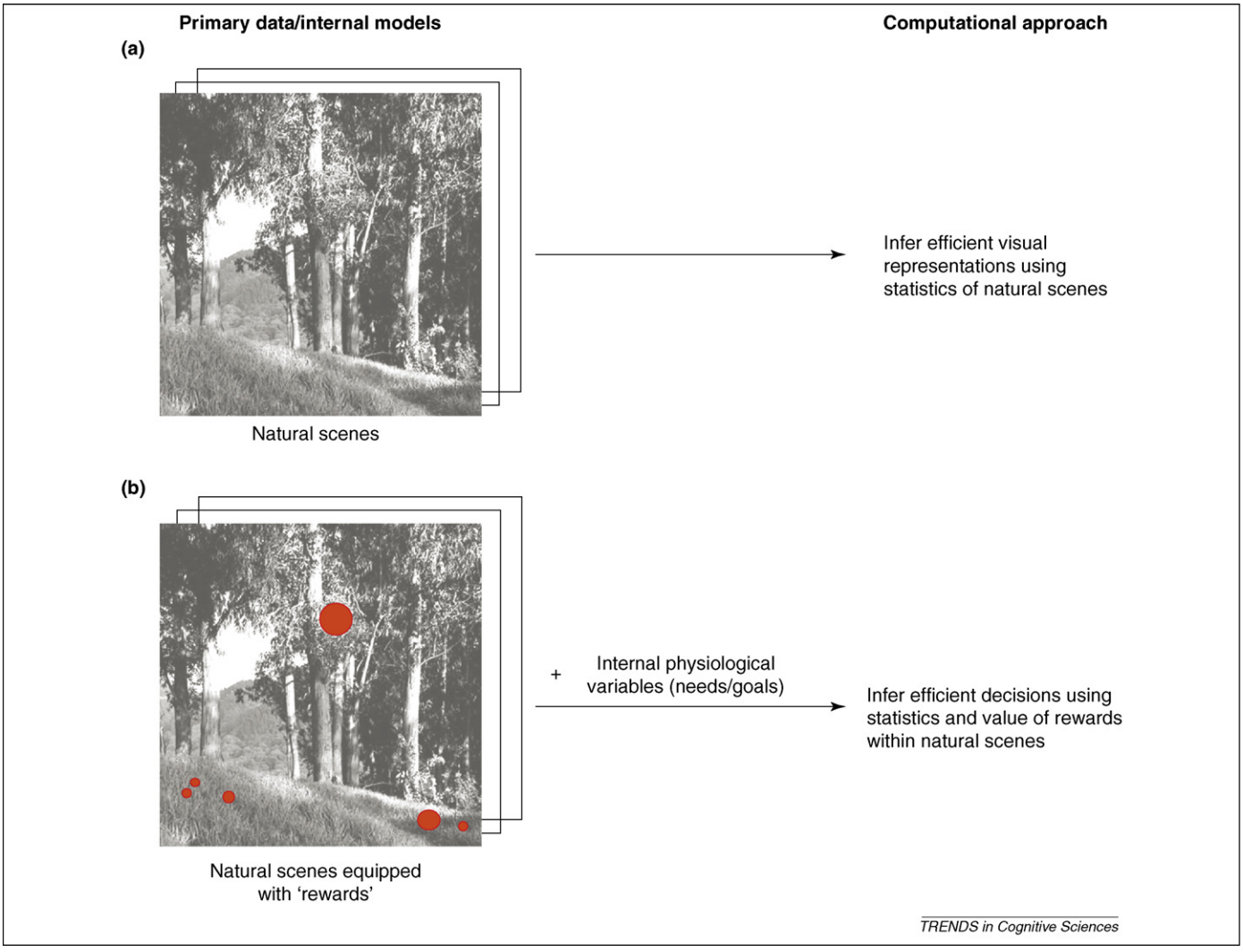

Figure 1. Defining the 'natural reward-harvesting statistics' problem. **(a)** The work of Horace Barlow crystallized an approach to vision by asking for the kinds of responses one should expect from visually sensitive neurons if they were efficient representations of 'natural visual statistics'. The natural visual statistics approach has expanded greatly in recent years to show that many aspects of the visual world are 'matched' in an efficient way by visual neural responses. Two features are particularly pertinent: first, local spatiotemporal correlations abound – the visual statistics from one point in visual space are highly correlated with neighboring points [12]; second, scale invariances also abound in natural visual statistics. **(b)** The reward-harvesting problem must have its own 'natural reward-harvesting statistics'. This kind of description of the reward acquisition problem would need to include at least two novel elements not naturally present in the visual problem and not usually included in optimal-foraging theory. The first is the need to consider the differential costs of exploring an environment to obtain the (possible) rewards present there. The second is the need to take account of the agent's state-of-motivation for the reward in question. Since rewards can be abstractly defined, it is now particularly important to generate good models of motivation.

probe neural and cognitive representations of visual experience ([10,11], see [12] for review). It is a vital and refreshing area of neuroscience because of its blend of theory, simulation and theory-guided experiments. Barlow himself has credited many other scientists with the same ideas [13]; however, we think it is fair to credit his ideas as an important starting point of the modern approaches to natural statistics in vision.

The natural statistics of reward-harvesting – matching signals from within and without

Beyond the natural visual statistics problem is an issue that all animals must solve to survive – they must be able to set and pursue goals, and one really important goal is to acquire food and water. Given our prelude to this point, a question naturally arises: Are there natural statistics of reward-harvesting? The short answer is of course that there is, but the proviso is that it is not simply a problem of reward distributions in the outside world. Instead, the problem also depends intimately on the current estimation of the many internal needs of the creature.

Were we to follow the approach described above for vision, this question would amount to asking whether there are efficient representations for reward-harvesting; that is, can we study the natural statistics of reward distributions in the world, discover redundancies in these distributions (i.e. redundancies in space, time and value), and then make educated guesses about the kinds of neural representations that would be efficient given these statistics? Big question. Part of the issue has already been addressed by the optimal-foraging literature, but not the full question. Let us unravel these issues a bit. First, let us consider in more detail what efficient representation means.

Exactly what is meant by efficiency in the context of neural representations? Efficiency in a representation is

equivalent to matching an encoding strategy to the natural statistical structure of the input 'signals'. The crucial words in this description are "matching" and "signals". As we outlined briefly above, matching for vision means that the visual system should have adopted processing strategies that reduce redundancy present in the visual inputs experienced over time. However, there is some vagueness in this account. One source of vagueness arises by not specifying the timescale(s) over which the matching process should happen. Some of the best work on natural visual statistics has addressed evolutionary timescales by examining visual neuron response properties that appear to be matched to statistical regularities present in *all* visual scenes [4,9,11]. These approaches address problems that must be solved by all creatures that use vision. But visual scenes change all the time for all kinds of reasons, including the movement of the creature; consequently, the matching mechanisms must be able to deal with a rather dramatic range of timescales (e.g. [14], also see [12] for review). To apply these ideas to reward-harvesting, one must identify clearly the source of the signals to which the creature's nervous system is adapted, and the most important timescales over which such signals fluctuate.

What is the best way to characterize the natural statistics of reward-harvesting? First, we must recognize that it represents a problem expanded from natural visual statistics, because the natural 'input' data originate from the outside world and from within the animal itself. As depicted in Figure 1b, the problem of wandering about and finding food day after day is difficult – consumables may be distributed haphazardly in space and time, they may run away using sophisticated escape strategies, and they may be of variable but difficult-to-judge quality. The list of difficulties is long. These signals arise 'from the outside', but a creature's body also generates an enormous array of 'inside' signals – both interoceptive and cognitive – that can change dramatically the value of the signals arriving from the outside world. Furthermore, these internal signals possess a natural statistics of their own, and that is the really difficult part of the problem.

Although it is relatively easy to point a light-sensitive camera at natural scenes and analyze the resulting images – moving or static – it is not clear what the 'natural interoceptive scene' would be. It is also not clear how one would easily measure such an internal scene moment-to-moment in order to subject it to the same kind of statistical analyses through which the visual world has been so nicely put [9–11]. Devising a general approach to measure natural statistics of 'internal' signals would represent a major step forward in this domain. Nevertheless, the

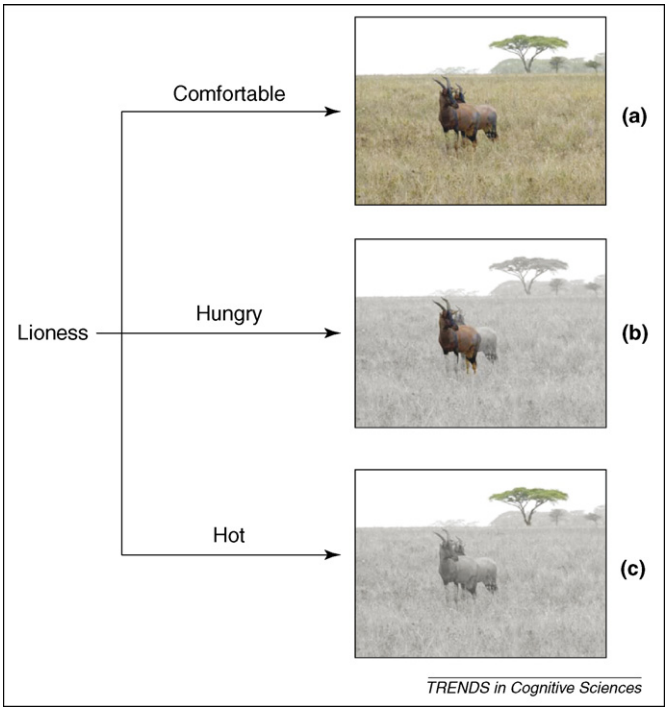

Figure 2. Matching interoceptive and perceptual representations. The internal states of an animal dramatically change the value of objects in its world. A hypothetical 'lioness' views a visual scene in three different states: **(a)** sated and comfortable; **(b)** hungry; and **(c)** sated but hot. These are not literal examples of how interoceptive states couple to visual perceptual representations, but they highlight the fact that there must be natural interoceptive statistics that have yet to be quantified, and the resulting representations should have an intimate connection to perceptual processing in the service of decision-making. Neuroeconomics, to be complete, must take on all these levels.

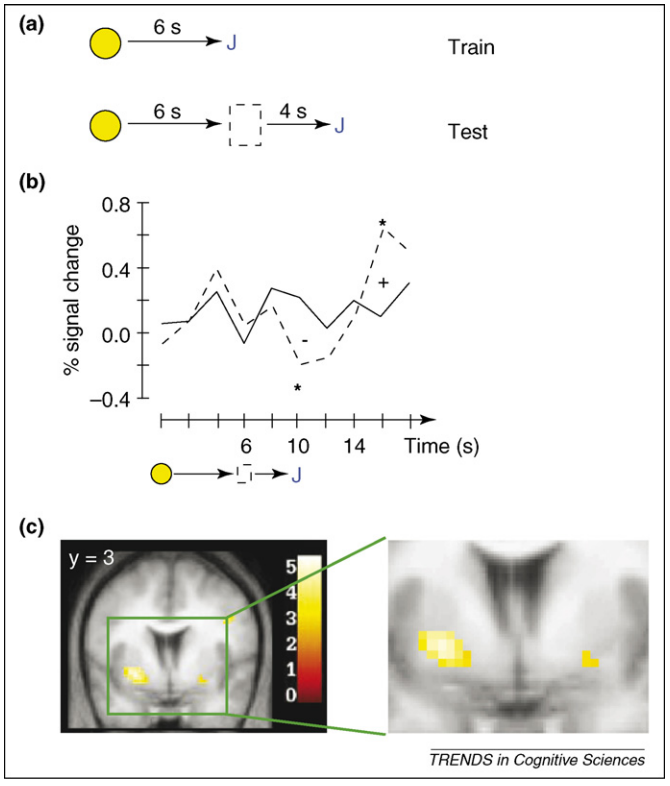

Figure 3. Hemodynamic signatures of learning signals during passive conditioning tasks. **(a)** During training, a squirt of juice (unconditioned stimulus) is delivered 6 s after the presentation of a yellow light (conditioned stimulus [CS]). Prior to training, the unpredicted delivery of the juice squirt elicits a positively signed prediction error, signifying that the animal has experienced something that is 'better than expected'. Following training, the delivery of the conditioned stimulus (light) elicits an expectation that juice will be delivered in 6 s. If the expected reward is withheld at 6 s following the light cue, a negatively signed prediction error signals that the animal has experienced something that was 'worse than expected'. However, if the withheld reward is later delivered at an unexpected delay (e.g. 10 s following the light cue), a positively signed prediction error is generated. **(b)** When humans undergo the procedure described above, the omission of an expected juice squirt (negative TD prediction error at 6 s post-CS) elicits a significant decrease in hemodynamic activity in the putamen, whereas subsequent delivery of the juice squirt at an unexpected delay (positive TD prediction error at 10 s post-CS) elicits a significant increase in hemodynamic activity (adapted from Ref. [37]). **(c)** Continuous prediction error signals generated throughout simple reward conditioning tasks similarly identify regions of the ventral putamen (adapted from Ref. [36]).

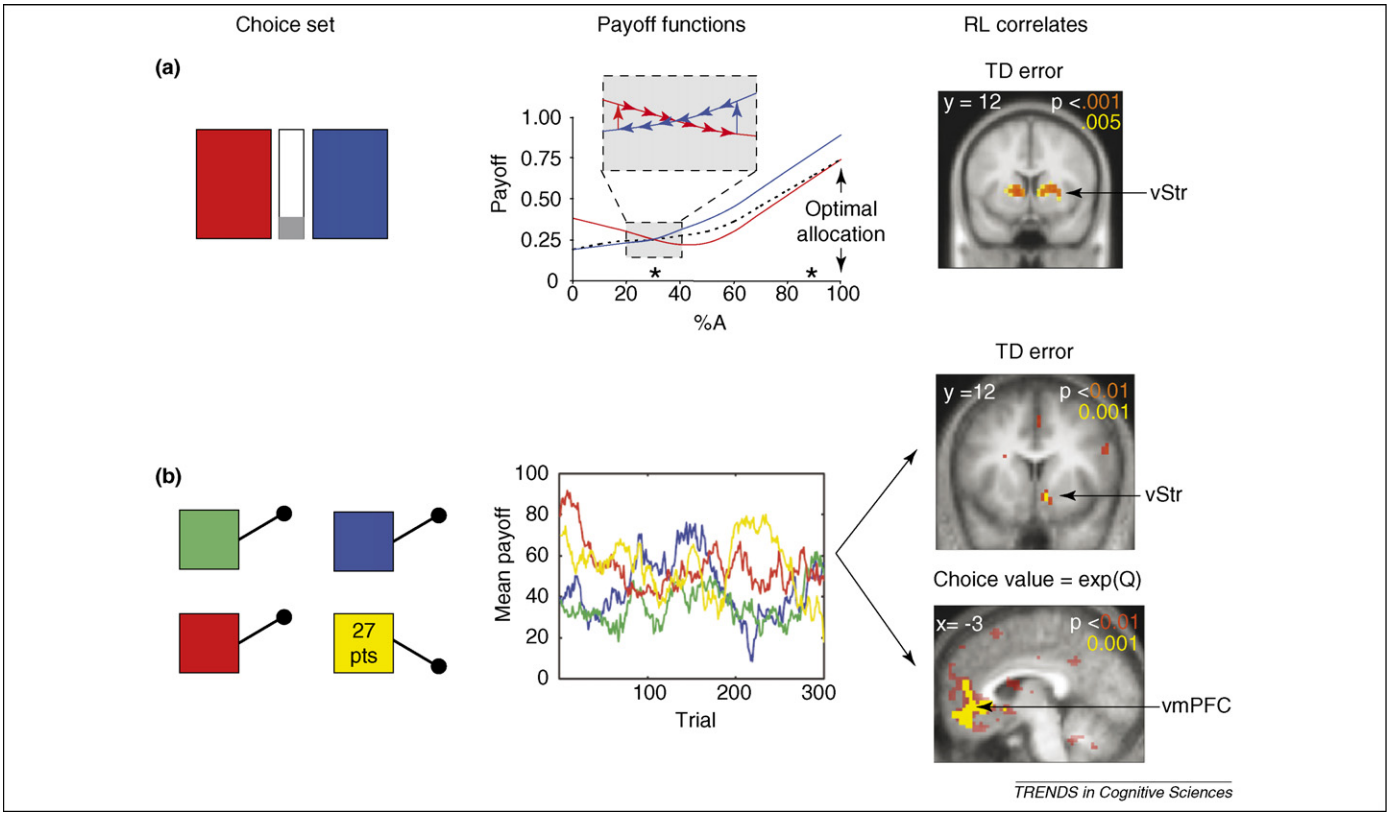

Figure 4. Hemodynamic signatures of learning signals during sequential decision-making tasks. **(a)** Subjects choose one of two actions (indicated by red or blue) within a sequential decision-making task and receive a monetary reward for each decision made (indicated by height of gray bar following action). The value of each choice was non-stationary, and was instead determined by the subject's own choice history. Specifically, the value of choosing red was determined by how often the subject had chosen red on the previous 20 trials. Thus, the red choice was worth more if the subject had previously chosen red 20% of the time than if the subject had previously chosen red 40% of the time. In fact, if the subject had previously chosen red 40% of the time, the value of choosing blue was greater than the value of red. Decisions within the task can be guided by prediction errors generated by the difference between estimates of the dynamically changing choice values and obtained reward. The continuously varying prediction error signal correlates with activity in the ventral striatum across 250 sequential choices (adapted from Ref. [39]). **(b)** Prediction errors associated with four choices in a sequential decision-making task. Within the 'four-arm bandit' task, the values of four options varied randomly and noisily, such that subjects were required to actively sample choices to maintain accurate estimates of choice values. Prediction errors elicited were scaled with response within the striatum, whereas value-related activation was identified within the ventromedial prefrontal cortex (adapted from Ref. [40]).

reward-harvesting problem has these two components – signals from within and without – both of which possess natural statistics.

Optimal foraging theory provides one approach to the 'outside' reward problem and has addressed internal constraints such as energetic state [15,16]. In this approach, reward-harvesting is posed as a problem of optimizing the net energy obtained by consuming some prey, wherein the idea of prey can range from mobile creatures that run away with sophisticated strategies to organisms that are immobile but provide variable amounts of net return (e.g. plants). One example of this approach is illustrated by the classic study of Krebs, Kacelnik and Taylor [17], who examined the nearly optimal foraging behavior of birds in a controlled setting. This same type of work on optimal foraging has been applied to starlings in cases in which their internal state (i.e. energetic state) was manipulated [18,19]. These studies raise the question of just how optimal the behavior of mobile creatures is, a question elegantly examined by Gallistel and colleagues [20] in their study of rats. They showed that a rat choosing between two alternatives on a variable-rate reward schedule approximates a Bayesian model of an ideal reward-rate detector!

Two summary points arise from the results highlighted above. First, the reward-harvesting problem has two general classes of signal – external and internal – but the internal class of signal has not yet been subjected to any serious statistical analysis that rivals the kind of work carried out on visual inputs [8–14]. Nevertheless, we should expect neural representations to handle efficiently both the inside and the outside signals. Second, in practical yet controlled settings, real creatures appear to be remarkably efficient at collecting rewards. Almost nothing is known about the nature of the representations animals use to match their choices to the reward-harvesting task demands arriving from their external and internal environments. Despite this great gap in our understanding of these representations, we do know that, whatever specific form they take, they have learning signals associated with them (Figures 2–4).

Connecting internal needs to perceptual representations

Modern efforts to understand internal representations for reward-harvesting in real creatures have gravitated to a style of learning model – called reinforcement learning (RL) models – that has been lifted in 'starter-kit' form from the machine-learning community [21,22]. Box 1 gives a brief sketch of this class of model. Although these models have helped frame the problem of reward-harvesting, they are

Box 1. The components of reinforcement learning models

Reinforcement learning (RL) models depict the learning problem for a reward-harvesting agent as an interaction among signals from the external world, internal representations, and teaching signals (learning signals) engendered by changes in the agent's internal state. The teaching signals possess an intimate relationship to the internal representations employed by the creature (as described below). In RL, as applied to real creatures, an animal's nervous system observes the outside world and its own internal state (both signals sources discussed above), produces an output (an action or change in internal state), and receives a scalar feedback signal that reflects the 'goodness' or 'badness' of the emitted action (i.e. it criticizes the action). The goal of learning for these models is to choose sequences of actions that maximize future-harvested reward.

Do we know anything about how the brain generates reinforcement signals capable of guiding (criticizing) the state changes experienced by an animal? The short answer is yes, but we begin by pointing out what an animal using RL must possess before encountering learning problems – representation, valuation and feedback signals. An animal using RL must possess (beforehand) a good depiction of the state space of the actual problem at hand (i.e. the representation piece); it must also begin with a modestly accurate value function defined over this representation (i.e. the valuation piece); and it must be able to generate an informative feedback signal using the representation and its associated value function. In biologically applied RL models, the value function (typically) associates with each state of the animal a number, its 'value', which represents the total reward that can be expected (on average) from that state into the distal future [22–25]. This kind of stored value is like a long-term judgment; it 'values' each state. It is these values that change through experience and under the guidance of the scalar feedback signal. The scalar feedback signal is called a 'reward error prediction signal' [21,22] and is analogous to a 'force' in the sense that it is the gradient of the value with respect to the state plus an extra part that is due to immediate unanticipated reward.

Reward prediction error (TD error) $= r\ (S_{t+1}) + \gamma V(S_{t+1}) - V(S_t)$ [Equation 1], where S_t is the current state and S_{t+1} is the next state to which the animal has transitioned. t can be pictured as time or simply a variable that provides a way to order the states visited. V is the value function that reflects the long-term value of each state, and r is the experienced reward (quoted from [24]; see [23,25,26]).

Equation 1 is the popular temporal difference (TD) error signal, which can be used directly to update the values of each state ([22]; see [23,25,27] for reviews). This framework is important to biologists because of its connection to identified neural systems, in particular the midbrain dopaminergic system. In the early 1990s, it was suggested that this reward prediction error signal was encoded by transients in midbrain dopamine neuron-firing rates in mammals and by similar transients in their octopaminergic analogues in insects [28–32]. At the time, the initial proposal provided a connection to learning algorithms with explicit supervisor signals (teaching signals) and with known convergence properties [33]. These connections to concrete biological systems and computational theory invited additional theory-building that connected RL models to neuromodulatory systems. The initial agreement of the theory and the data is quite provocative [34,35], but more importantly RL models now inform the design and interpretation of a range of experiments too numerous to detail here [27]. Here, we would like to focus on two aspects of these models: the abstract nature of the idea of a 'reward prediction error signal'; and the brain's capacity to use seemingly arbitrary information to define the 'reward signal' in Equation 1 above. These two features will show how RL systems implicitly define a common internal currency for use during valuation and decision-making (see Box 2).

Box 2. Reward prediction error signals and common currencies

Three basic components were described above for RL models:

(i) **Frame** the problem. Pick a representation of the state-space and 'reward' signal appropriate to the problem at hand, along with an associated valuation over the state-space.

(ii) **Generate** the 'reward prediction error signal' by taking action or changing internal state.

(iii) **Choose** the action. Use policy (behavioral strategy) that maps states to actions according to the value function over the state space.

Figure 3 illustrates two separate passive conditioning tasks in which a cue (light) predicted the later delivery of a juice squirt. One experiment tracked a reward prediction error signal during training [36], and the other group over-trained subjects on specific timing and inserted catch trials to induce large negative and positive prediction errors [37]. The groups found these reward prediction errors in overlapping regions in the putamen [36,37]. The heavy dopaminergic projections into this region, and the electrophysiological results in monkeys make these findings an expected result [35]; however, reward prediction errors can also be tracked in higher-order conditioning tasks in humans and by using an aversive outcome for the 'reward' signal [38].

Reward prediction errors can also be tracked in sequential decision tasks and used as inputs to a policy that chooses actions (Figure 4) [39,40]. This modeling approach matches well the sequence of choices made by the human subjects and provides for each subject a customized reward prediction error signal, which shows up as before in the striatum (Figure 4). Brain responses reflecting the value function can also be identified (Figure 4b) [40].

The literature now contains many more examples that make these same points, but such examples are too numerous to detail here. However, we propose that the impact of reward prediction error signals on behavior (on the choice tasks) along with their prominent neural correlates qualifies them as a kind of 'differential neural currency'. Such differential currencies are ideally suited to compare the relative value of different behavioral options available to each subject given the task at hand.

The concept of 'reward signal' is quite general and not limited to primary rewards such as food and water, but it appears to extend to nearly any idea that can be conjured in a human brain. Second, although we know almost nothing about the exact nature of the representations that underlie human behavior in response to these tasks, these experiments, seen through the filter of RL, show that these representations – presumably different from one another – can generate signals well captured by the dynamics of reward prediction errors. Moreover, the subjects' nervous systems appear to be quite willing to choose on the basis of these reward prediction errors, which in our opinion qualifies the error signal as a generic 'differential currency'. We use the term ''differential'' to emphasize that the evidence best supports a case of relative valuation. This way of framing reward prediction errors fits well with the notion of dopamine bonus championed by Kakade and Dayan [41] to explain a variety of so-called anomalous responses in dopaminergic neurons not explained easily as a reward prediction error signal. Their language is exactly one that depicts extra dopamine transients as a kind of extra currency, a view also supported by the computational model of addiction suggested by Reddish [42].

really only semi-quantitative at this point in their application to biological systems; that is, they prescribe the style of learning signals to expect in biological systems, the polarity of these signals in specific circumstances, and a few relevant parameters (e.g. learning rate and discount rate) that can be extracted from experiments. Underlying all the reinforcement-learning models is the goal of learning a valuation function for the problem at hand (see Box 2). This function provides the adaptive creature with a way to value its internal states and its ongoing sensory experience. The 'semi' in our depiction of these models as semi-quantitative relates specifically to the lack of understanding of the representations used by real creatures to pursue real rewards – whether they be primary rewards, such as food

and water, or more abstract rewards, such as 'earning more money' or 'having a better home life'.

Conclusions

In closing, we should point out that efficient statistical approaches to sensory information – however informative they may be when portrayed 'open-loop' (Figure 1) – will be incomplete until they can be paired with a related treatment of the internal states and needs of the organism. As illustrated in Figure 2, this pairing may be profound. Our proposal is that interoceptive representations have a close and intimate connection to perceptual representations, with the idea of efficiency tying the two together. A sated and comfortable lioness looking at two antelopes sees two unthreatening creatures against the normal backdrop of the temperate savanna (Figure 2a). The same lioness, when hungry, sees only one thing – the most immediate prey (Figure 2b). In another circumstance, in which the lioness may be inordinately hot, the distant, shaded tree becomes the prominent visual object in the field of view (Figure 2c); the mismatch between the internal need (to stay at comfortable temperature) and the external signals (it is hot outside) changes the importance of the visual signals. In this context, it is not altogether clear how to define 'efficient'; however, we suspect that such efficiencies have already been discovered, even though they have not been defined by our descriptions of interoception or sensory perception.

These illustrations are not meant literally but make the point that efficiency in visual representations cannot ignore the internal state of the animal, and so one should expect central representations of interoceptive states and perceptual representations (in this case for vision) to be closely allied if not part of the same neural information streams. With the acceleration of non-invasive imaging technologies, we suspect that this internal natural statistical problem will become tractable.

References

1 Glimcher, P.W. (2003) *Decisions, Uncertainty, and the Brain. The Science of Neuroeconomics,* MIT Press
2 Montague, P.R. and Berns, G.S. (2002) Neural economics and the biological substrates of valuation. *Neuron* 36, 265–284
3 Bialek, W. (1987) Physical limits to sensation and perception. *Annu. Rev. Biophys. Biophys. Chem.* 16, 455–478
4 Laughlin, S.B. (1994) Matching coding, circuits, cells, and molecules to signals. General principles of retinal design in the fly's eye. *Prog. Retinal Eye Res.* 13, 165–196
5 van Beers, R.J. *et al.* (2002) 2002. Role of uncertainty in sensorimotor control. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 357, 1137–1145
6 MacKay, D. and McCulloch, W. (1952) The limiting information capacity of a neuronal link. *Bull. Math. Biophys.* 14, 127–135
7 Barlow, H.B. (1961) Possible principles underlying the transformation of sensory messages. In *Sensory Communication* (Rosenblith, W., ed.), pp. 217–234, MIT Press
8 Barlow, H.B. and Foldiak, P. (1989) Adaptation and decorrelation in the cortex. In *The Computing Neuron* (Miall, C. *et al.*, eds), pp. 54–72, Addision-Wesley
9 Atick, J.J. (1992) Could information-theory provide an ecological theory of sensory processing? *Network* 3, 213–251
10 Ruderman, D.L. (1994) The statistics of natural images. *Network* 5, 517–548
11 Olshausen, B.A. and Field, D.J. (2004) Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487
12 Simoncelli, E. and Olshausen, B.A. (2001) Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24, 1193–1215
13 Barlow, H.B. (2001) Redundancy reduction revisited. *Network* 12, 241–253
14 Brenner, N. *et al.* (2000) Adaptive rescaling maximizes information transmission. *Neuron* 26, 695–702
15 Kamil, A.C. *et al.* (1987) *Foraging Behavior,* Plenum Press
16 Maynard Smith, J. (1982) *Evolution and the Theory of Games,* Cambridge University Press
17 Krebs, J.R. *et al.* (1978) Tests of optimal sampling by foraging great tits. *Nature* 275, 27–31
18 Marsh, B. *et al.* (2004) Energetic state during learning affects foraging choices in starlings. *Behav. Ecol.* 15, 396–399
19 Pompilio, L. and Kacelnik, A. (2005) State-dependent learning and suboptimal choice: when starlings prefer long over short delays to food. *Anim. Behav.* 70, 571–578
20 Gallistel, C.R. *et al.* (2001) The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. *J. Exp. Psychol. Anim. Behav. Process.* 27, 354–372
21 Sutton, R.S. (1988) Learning to predict by the methods of temporal difference. *Mach. Learn.* 3, 9–44
22 Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning,* MIT Press
23 Montague, P.R. *et al.* (2004) Computational roles for dopamine in behavioural control. *Nature* 431, 760–767
24 Montague, P.R. *et al.* (2006) Imaging valuation models in human choice. *Annu. Rev. Neurosci.* 29, 417–448
25 Doya, K. (2002) Metalearning and neuromodulation. *Neural Netw.* 15, 495–506
26 Daw, N.D. *et al.* (2005) Actions, policies, values, and the basal ganglia. In *Recent Breakthroughs in Basal Ganglia Research* (Bezard, E., ed.), pp. 91–106, Nova Science Publishers
27 Daw, N.D. and Doya, K. (2006) The computational neurobiology of learning and reward. *Curr. Opin. Neurobiol.* 16, 199–204
28 Quartz, S. *et al.* (1992) Expectation learning in the brain using diffuse ascending connections. *Soc. Neurosci. Abst.* 18, 1210
29 Montague, P.R. *et al.* (1993) Using aperiodic reinforcement for directed self-organization. *Adv. Neural Inf. Process. Syst.* 5, 969–976
30 Montague, P.R. *et al.* (1995) Bee foraging in an uncertain environment using predictive Hebbian learning. *Nature* 376, 725–728
31 Sejnowski, T.J. *et al.* (1995) Predictive Hebbian learning. *COLT* 8, 15–18
32 Montague, P.R. *et al.* (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* 16, 1936–1947
33 Dayan, P. (1992) The convergence of TD(λ) for general λ. *Mach. Learn.* 8, 341–362
34 Schultz, W. *et al.* (1997) A neural substrate of prediction and reward. *Science* 275, 1593–1599
35 Schultz, W. (1998) Predictive reward signals of dopamine neurons. *J. Neurophysiol.* 80, 1–27
36 O'Doherty, J.P. *et al.* (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38, 329–337
37 McClure, S.M. *et al.* (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38, 339–346
38 Seymour, B. *et al.* (2004) Temporal difference models describe higher order learning in humans. *Nature* 429, 664–667
39 Li, J. *et al.* (2006) Policy adjustment in a dynamic economic game. *PLoS ONE* 1, e103
40 Daw, N.D. *et al.* (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441, 876–879
41 Kakade, S. and Dayan, P. (2000) Dopamine bonuses. *NIPS* 2000, 131–137
42 Redish, A.D. (2004) Addiction as a computational process gone awry. *Science* 306, 1944–1947