From the Strüngmann Forum Report: Better Than Conscious?
Decision Making, the Human Mind, and Implications For Institutions
Edited by Christoph Engel and Wolf Singer
 ISBN-10: 0-262-19580-1 ISBN-13: 978-0-262-19580-5
http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=11498

3

The Role of Value Systems in Decision Making

Peter Dayan

Gatsby Computational Neuroscience Unit, UCL, London WC1N 3AR, U.K.

Abstract

Values, rewards, and costs play a central role in economic, statistical, and psychological notions of decision making. They also have surprisingly direct neural realizations. This chapter discusses the ways in which different value systems interact with different decision-making systems to fashion and shape affectively appropriate behavior in complex environments. Forms of deliberative and automatic decision making are interpreted as sharing a common purpose rather than serving different or non-normative goals.

Introduction

There is perhaps no more critical factor for the survival of an organism than the manner in which it chooses between different courses of action or inaction. A seemingly obvious way to formalize choice is to evaluate the predicted costs and benefits of each option and pick the best. However, seething beneath the surface of this bland dictate lies a host of questions about such things as a common currency with which to capture the costs and benefits, the different mechanisms by which these predictions may be made, the different information that predictors might use to assess the costs and benefits, the possibility of choosing when or how quickly to act as well as what to do, and different prior expectations that may be brought to bear in that vast majority of cases when aspects of the problem remain uncertain.

In keeping with the complexity and centrality of value-based choice, quite a number of psychologically and neurally different systems are involved. These systems interact both cooperatively and competitively. In this chapter, I outline the apparent dependence of four decision-making mechanisms on four different value systems. Although this complexity might seem daunting, we will see that exact parallels can be found in ideas about how artificial systems such as

robots might make choices, that the different systems capture rather natural trade-offs that are created by the statistical and computational complexities of optimal control, and that at least some of the apparent problems of choice actually arise from cases in which reasonable a priori expectations about the world are violated by particular experimental protocols.

I begin by describing a formal framework that derives from the field of reinforcement learning (Sutton and Barto 1998). Reinforcement learning mostly considers how artificial systems can learn appropriate actions in the face of rewards and punishments. For our purposes, however, it is convenient since it (a) originated in mathematical abstractions of psychological data, (b) has a proof-theoretic link to statistical and normative accounts of optimal control, and (c) lies at the heart of widespread interpretations of neurophysiological and neuroimaging data. I show the roles played by value-based predictions in reinforcement learning and indicate how different psychological and neurobiological ideas about decision making map onto roughly equivalent reinforcement learning notions.

Thereafter, I describe four different value systems and four different decision-making systems that arise naturally within this framework. Two of the decision makers—one associated with goal-directed or forward model-based control, the other associated with habitual control (Dickinson and Balleine 2001)—are well characterized neurally; the substrates of the other two are somewhat less clear. Equally, the anatomical (e.g., neuromodulatory) bases of two of the value systems are somewhat clearer than those of the other two. Different decision makers use value information directly and through learning in different ways; they also enjoy cooperative and competitive relations with each other. The extent to which these interactions might underlie some of the complexity in the data on choice is then considered. I conclude with a discussion of the key open issues and questions.

A few remarks at the outset: First, we would like to aim toward a theory of behavior, not just of "decisions," somehow more narrowly defined. That the systems involved in normative choice appear also to be implicated in the tics of a patient with Tourette's syndrome or the compulsions of one with obsessions suggest that we should not start by imposing arbitrary boundaries. Second, our analysis owes an important debt to Dickinson (e.g., 1994), who, adopting a view that originated with Konorski (1948, 1967), and together with his colleagues (e.g., Dickinson and Balleine 2001), has long worked on teasing apart the various contributions from different systems. Third, we will typically lump together information derived from rodent, monkey, and human studies. We are far from having a sufficiently refined understanding to be able sensibly to embrace the obviously large differences between these species. Finally, the limit on the number of citations has forced the omission of many highly relevant studies.

Reinforcement Learning and Dynamic Programming

The problem for decision making can be illustrated in the simple maze-like task in Figure 3.1. It is helpful to define a formalism which allows us to describe the external state of the subject ($x \in \{x_1, x_2, x_3\}$, its position in the maze), its internal state (m, mostly motivational factors such as hunger or thirst), its possible choices ($c \in \{L, R\}$), and a set of possible outcomes ($o \in \Omega =$ {cheese, nothing, water, carrots}, but in general also including aversive outcomes such as shocks).

Immediate Outcomes

Let us first consider the case that the subject has just one choice to make at a single state, for instance x_3 in the maze. We write the probability of receiving outcome o given choice c as $O_{co}(x)$. This would be stochastic if, for instance, the experimenter probabilistically swaps the outcomes. To select which action to execute, the subject must have preferences between the outcomes. One important way to describe these preferences is through a utility function $u_o(m)$, which we call a native (or experienced) utility function (Kahneman et al. 1997). This should depend on the motivational state of the subject since, for instance, all else being equal, thirsty subjects will prefer the water, and hungry subjects, the cheese. Example utilities for all the outcomes are shown in Figure 3.1b.

The native utility function quantifies the actual worth to the subject of the outcome. It is the first of the four value functions that we will consider overall. Given the utility, the subject could choose normatively the action that maximizes the expected utility of the choice, or, more formally, defining these as the Q values of choice c at environmental and motivational states x and m, respectively,

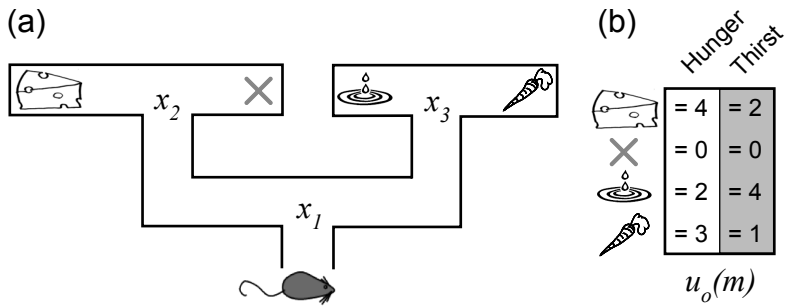

Figure 3.1 Maze task: (a) A simple maze task for a rat, comprising three choice points (x_1, x_2, x_3) and four outcomes ($o \in$ {cheese, nothing, water, and carrots}). The rat has to run forward through the maze and must go left (L) or right (R) at each choice point. (b) Utilities $u_o(m)$ for the four outcomes are shown for the two motivational states, m = hunger and m = thirst. Figure adapted from Niv, Joel, and Dayan (2006).

$$\mathcal{Q}_c(x,m) = \sum_o O_{co}(x) u_o(m), \tag{1}$$

it could choose:

$$c^*(x,m) = \text{argmax}_c \left\{ \mathcal{Q}_c(x,m) \right\}. \tag{2}$$

The systematic assignment of choices to states is usually called a *policy*, a term which comes from engineering. Policies can be deterministic (as here) or, as is often found experimentally, probabilistic, with the subject choosing all actions, but some (hopefully, the better ones) more frequently than others. One conventional abstraction of this stochastic choice is that the probability of executing action c is determined by competition between $\mathcal{Q}$ values, as in a form of Luce choice rule (Luce 1959) or softmax:

$$P(c;x,m) = \frac{\exp\left(\beta \mathcal{Q}_c(x,m)\right)}{\sum_d \exp\left(\beta \mathcal{Q}_d(x,m)\right)}. \tag{3}$$

Here, parameter $\beta > 0$ (sometimes called an inverse temperature) regulates the strength of the competition. If β is small, then choices that are worth very different amounts will be executed almost equally as often.

Straightforward as this may seem, there are conceptually quite different ways of realizing and learning such policies. These different methods are associated with different combinations of value and decision-making systems. Key combinations are described briefly here; the separate systems and their interactions will be discussed later.

1. Perhaps the simplest approach would be to implement Equations 1 and 2 directly by learning a so-called *forward model* $\hat{O}$ [note: hats are used to indicate estimated, approximate, or learned quantities]; that is, the consequences of each action, together with a way of estimating the utility of each outcome, $\hat{u}$. If the contributions can be accumulated, as in Equation 1, then each action could compete according to its estimated $\hat{\mathcal{Q}}$ value (a form of predicted utility; Kahneman et al. 1997). Figure 3.2a shows how this works for the case of state x_3 in the maze. The neural substrate of this means of evaluating the worth of options is a second value system influencing choice. This will be related below to the psychological notion of a goal-directed controller implemented in the prefrontal cortex and dorsomedial striatum.
2. One alternative to the computational and representational costs of this scheme would be to try and learn the $\mathcal{Q}$ values directly and then only have those compete. This would obviate the need for learning and using O. Figure 3.2b shows these values, again for the case of x_3.

   A mainstay of psychology, the Rescorla-Wagner learning rule, which is equivalent to engineering's delta rule, suggests one way of doing this.

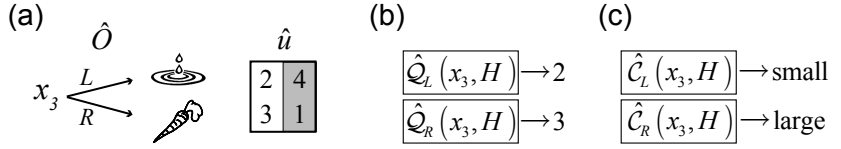

Figure 3.2 Decision making at x_3: (a) The forward model consists of $\hat{O}$ and $\hat{u}$. By exploring the explicit options, the subject can work out which action is best under which circumstance. (b) The $\hat{\mathcal{Q}}$ values suffice to indicate the expected utilities for each action under a motivational state (in this case, hunger), but the provenance of these values in the actual outcomes themselves is not accessible. (c) An even more direct controller just specifies ordinal quantities $\hat{C}$ on which choice can be based; these numbers, however, lack anything like an invertible relationship with expected utilities. Figure adapted from Niv, Joel, and Dayan (2006).

According to this rule, the estimate $\hat{\mathcal{Q}}_c(x, m)$ is represented, perhaps using a set of synaptic weights, and when action c is tried, leading to outcome o, the estimate is changed according to the prediction error:

$$\hat{\mathcal{Q}}'_c(x,m) = \hat{\mathcal{Q}}_c(x,m) + \alpha\left[u_o(m) - \hat{\mathcal{Q}}_c(x,m)\right], \tag{4}$$

where α is called a learning rate (or associability). This rule has a normative basis and will often lead to $\hat{\mathcal{Q}} = \mathcal{Q}$, at least on average. There is good evidence that a prediction error closely related to this is represented in the activity of dopamine neurons (Montague et al. 1996; Schultz 1998).

The structure that expresses estimates like $\hat{\mathcal{Q}}_c(x, m)$ is actually the third of the value systems that plays a role in decision making. Such values are sometimes called *cached* (Daw et al. 2005) because they cache information actually observed about future outcomes in the form of expected utilities.

The nature of the dependence on the motivational state m is a critical facet of cached values (see Dayan and Balleine 2002). In general, there is ample evidence for state dependence in learning and recall (i.e., m might indeed be involved in the representation of the $\hat{\mathcal{Q}}$ values). However, consider the case that a subject learns $\hat{\mathcal{Q}}$ values in one motivational state (e.g., hunger, as in Figure 3.2b). Since these values are simply numbers, utilities that are divorced from their bases in actual outcomes, they can be expected to generalize promiscuously, certainly more so than values from the forward model. In particular, if they generalize to other motivational states such as thirst, then the subjects may continue to favor actions leading to food, even though this is inappropriate. To put it another way, there may be no way to move from the $\mathcal{Q}$ values appropriate to hunger to the ones suitable for thirst without explicit relearning. This type of inflexibility is indeed a hallmark of behavioral habits, as we explore in more detail below.

3. An even simpler scheme would be to use a choice mechanism such as that in Equation 2, but to observe that the policy only requires that the

value used in the argmax associated with action $c^*(m, x)$ has to be numerically larger than the values associated with all the other actions, and not that it actually satisfies Equation 1. Figure 3.2c demonstrates this, using $\hat{C}_c(x, m)$ for the action values. Learning rules for such values are closely related to the psychological notion of stimulus-response learning, with a choice being stamped in or reinforced by large delivered utilities. We treat these action values as also being cached, since they share critical characteristics with the $\mathcal{Q}$ values (e.g., they are only indirectly sensitive to motivational manipulations).

In short, there are different mechanisms, all of which can be used to achieve the same ostensive goal of optimizing the net expected value. The differences will become apparent when aspects of the motivational state that pertained during training are different from those during the test. Understanding how the different systems generalize is therefore critical. Bayesian ideas are rapidly gaining traction as providing general theories of this, encouraging careful consideration of prior expectations, here in the nature of any or all of O, u, $\mathcal{Q}$, and c as well as the internal representations of x and m underlying these functions. I will argue that such priors play a central role in everything from Pavlovian conditioning (arising from a prior over actions appropriate to predictions of reward or punishment) to learned helplessness (a prior over the nature of the controllability of an environment).

Delayed Outcomes

One important characteristic of many choice problems is that the outcome may not be provided until a whole sequence of actions has been executed. In the maze shown in Figure 3.1, this is true for state x_1, as either action only gets the subject to x_2 or x_3, and not directly to any of the outcomes. In fact, many outcomes may be provided along a trajectory and not just a single one at the end. In simple (Markovian) environments like the maze, it is possible to formalize trajectories by considering transitions from one state x to others y whose probabilities $T_{xy}(c)$ may depend on the action chosen c.

A key definitional problem that arises is assessing the worth of a delayed outcome. Various schemes for doing so have been suggested in economics and psychology. One idea is *discounting*, with an outcome o that will be received in the future after time t, having a reduced net present value of $u_o(m) \times f(t)$, where $f(t) = \exp(-\gamma t)$ might be exponential or $f(t) = 1/(1 + \gamma t)$ hyperbolic ($\gamma > 0$). Under discounting, it becomes advantageous to advance outcomes with positive utilities and delay those with negative utilities.

Exponential discounting is akin to using a form of interest rate. It turns out that there are natural extensions to the definition of $\mathcal{Q}$ values that are suitable for trajectories (Watkins 1989), and necessitating only a small change to the

learning rule in Equation 4. Indeed, much of the data about the neural basis of value systems depends on these extensions.

Although hyperbolic discounting generally fits behavioral data more proficiently, it is not clear how this might be implemented without imposing implausible requirements on memory for learning. Hyperbolic discounting infamously leads to temporal inconsistencies (see Ainslie 2001). For instance, if $\gamma = 1\ s^{-1}$, then although a reward of 4 units that arrives in 10 s would be preferred to one of 3 units in 9 s, after 8 seconds has passed, the reward of 3 units, which would now arrive in 1 s, would be preferred to the reward of 4 units, which would arrive in a further 2 s. The latter pattern, in which a larger, but later reward is rejected in favor of a smaller, but earlier one, is called *impulsive*.

An alternative method for handling delayed outcomes is to optimize the *rates* at which rewards are received and costs avoided. Most of the consequences of this are too complex to describe here; however, in this scheme a key role is played by the fourth value system, which evaluates and predicts the long-run average rate (ρ) of the delivery of utility. The rate is the sum of all the utility received over a long period of time (from all sources), divided by that time period. The rate turns out to matter most when we consider optimizing the choice of not just *which* action, but also its alacrity, speed or vigor, given that the cost of acting quickly is taken into account (Niv, Daw et al. 2006; Niv, Joel, and Dayan 2006).

Four Value Systems

Having introduced the four value systems above, I will now summarize some of their properties. Note, in particular, the distinction between the true utilities $u_o(m)$ and true $\mathcal{Q}$ values $\mathcal{Q}_c(x, m)$ that depend on them, and the various estimates (distinguished using hats) which depend in different ways on the various underlying learned quantities. As will be apparent, many uncertainties remain.

Native Values

The actual utility (or experienced utility, in terms of Kahneman et al. 1997) $u_o(m)$ of an outcome in a motivational state m (or at least the most immediate neural report of this) should ground all methods of choice. Berridge and Robinson (1998) discuss extensive evidence about the existence of a so-called *liking* system, which reports on the motivational state-dependent worth of an outcome. Although its exact anatomical substrates are not completely clear, for food reward, liking has been suggested as being mediated by structures such as the primary taste system in the brainstem; further, neurons in the hypothalamus, for instance, have the ability to sense certain aspects of the internal state of an animal (e.g., hydration), and thus could play an important role. Nuclei in the pain system, such as the periacqueductal gray, may play a similar role

in the mediation of primary aversion (i.e., disliking). Opioids and benzodiazepines can act to boost liking and reduce disliking, perhaps by manipulating this value.

Forward-model Values

The second value system involves learning and using a forward model of the probabilities $\hat{O}$ of the outcomes (and, in the case of trajectories, the transitions $\hat{T}$) consequent on a choice, together with the motivational state-dependent utility $\hat{u}$ of those outcomes (Figure 3.2a). This involves the use of a model in the service of computing a predicted utility (Kahneman et al. 1997). There is evidence in various species that prefrontal areas (notably dorsolateral prefrontal cortex, the dorsomedial striatum, and the insular cortex) are involved in this sort of model-dependent predictions of future outcomes (and future states). The nature and substrate of the assessment of the state-dependent utility $\hat{u}$ is less clear, and various suggestions have been made about the psychological and neural mechanisms involved. For instance, one idea involves *incentive learning* (Dickinson and Balleine 2001); namely, that subjects learn the utilities from direct experience and the observed native utility. Another notion is that of a *somatic marker*; crudely that the body itself (or a cortical simulacrum thereof) is used to evaluate the native utility of a predicted outcome via a mechanism of top-down influence (Damasio et al. 1991).

One interesting question for these schemes concerns the motivational state m used to make the assessments. If this is the subject's current motivational state, but the prediction is about an outcome that will be obtained in a different motivational state, then the prediction might be wrong. In discussing "hot" and "cold" cognition, Loewenstein (1996) makes just such a point.

Cached Values

The $\mathcal{Q}$ values (Watkins 1989) discussed above and shown in Figure 3.2b can be seen as alternative estimates of predicted utilities (Kahneman et al. 1997). Unlike the native or forward model-based values, these $\mathcal{Q}$ values pay no direct allegiance to their provenance in terms of particular outcomes. This turns out to have distinct computational benefits for choice, particularly in the case of optimizing actions over whole trajectories. However, as has already been discussed, it has problems coping efficiently with changes to the environment or motivational state.

In the trajectory case, there is an important role for the cached value $V(x, m)$ of the state x in motivational state m, which is defined by averaging the choices over the policy, as either

$$V(x,m) = \mathcal{Q}_{c^*}(x,m) \qquad \text{given Equation 2, or} \qquad (5)$$

$$= \sum_c P(c; x, m) \mathcal{Q}_c(x, m) \text{ given Equation 3.} \quad (6)$$

Estimates $\hat{V}$ can be learned using a learning rule similar to that in Equation 4.

There is strong evidence for the involvement of the phasic release of the neuromodulator dopamine in the appetitive aspects of learning these cached values (e.g., Montague et al. 1996; Schultz 1998), and for the basolateral nucleus of the amygdala, the orbitofrontal cortex, and regions of the striatum in representing the $\hat{V}$ and $\hat{\mathcal{Q}}$ values. This has all been reviewed extensively elsewhere (e.g., O'Doherty 2004; Balleine and Killcross 2006), including in its ascription by Berridge and Robinson (1998) to the second of their two systems influencing choice ("wanting," which complements liking). The reader is referred to these excellent sources.

There is rather less agreement about how the aversive components of cached utilities are represented and manipulated, although there are ample data on the involvement of the amygdala and insula. An important psychological idea is that of opponency, between appetitive and aversive cached value systems. Indeed, some of the best psychological evidence for the existence of outcome-independent value systems, such as $\mathcal{Q}$ values, comes from facets of opponency. For instance, the *non*delivery of an expected *shock* can play some of the critical roles through learning of the delivery of an *un*expected *reward* (see Dickinson and Balleine 2001). It has been suggested that another neuromodulator, serotonin, has a starring role in the cached aversive system (Daw et al. 2002), but most of the evidence is rather circumstantial.

There may be a critical anatomical and functional separation between the systems associated with representing the cached values of states $\hat{V}$ and those representing the cached $\hat{\mathcal{Q}}$ values (O'Doherty 2004). This would be particularly important if cached action values like those considered in Figure 3.2c are employed, as these are further divorced from the utilities.

Long-run Average Value

The final notion of value is the long-run average utility ρ. This is important when it is the average utility per unit time that must be optimized, which is indeed the case for large classes of psychological experiments in such things as free operant schedules, for which rates of responding (i.e., choices about the times at which to act) are more central than choices between punctate actions. In the context of the maze, this might translate to the speed of running rather than the choice of which direction to run.

The importance of the long-run average value in controlling response rates has only fairly recently been highlighted, notably in the work of Niv, Daw et al. (2006) and Niv, Joel and Dyan (2006). They point out the link between ρ and the opportunity cost for behaving slothfully. When ρ is big and positive, animals should act faster, since every idle moment is associated directly with

a greater amount of lost (expected) utility. This helps explain why animals that are highly motivated by the expectation of receiving large rewards per unit time perform all actions quickly (and often not just those actions that lead directly to the reward).

In principle, ρ could be realized in either a forward model-based or a cache-based manner. However, Niv, Daw et al. (2006) suggest that the existing evidence, again for rewards rather than punishments, may favor an assignment of ρ to the tonic activation of dopaminergic neurons.

Again, less work has been done on the nature and realization of long-run average negative values. It is notable that various forms of depression have been postulated as depending in some way on stress and excessive affective negativity. We might expect these to be associated with a form of anergia (i.e., actions slowed are expected punishments postponed). However, more sophisticated effects of prior expectations appear to be at work in other paradigms used to capture aspects of depression, such as learned helplessness.

Four Decision Makers

As outlined above, these four value systems contribute in various ways to the different structures involved in decision making. Indeed, most of the evidence about the value systems derives from observations of choices that are inferred to depend on values of one sort or another. In this section, I briefly describe first the decision-making systems, and then some critical aspects of their interaction.

Goal-directed Control

The forward model is the most straightforward control mechanism. It was discussed in the description of Figure 3.2a. Choices are based on approximate evaluations ($\hat{u}$) of predicted ($\hat{O}$) outcomes. If the evaluation system $\hat{u}$ is sensitive to motivational state, then so will be the choice of action.

Figure 3.3a shows the extension of Figure 3.2a to the case of the full maze. The main difference is that the search process starts at x_1 and includes the use of $\hat{T}$ to work out which states follow which actions until outcomes are reached. By searching to the end of the tree, the subject can favor going left at x_1 when hungry, but right at x_1 when thirsty. This motivational dependence of control, which itself is a function of the motivational dependence of $\hat{u}$, is why this controller is typically deemed to be *goal-directed*. Since choices are based on search through the tree, which may be considered to be working-memory intensive and deliberative (see Sanfey et al. 2006), this controller distributes information in exactly the correct way. For instance, if the subject is hungry, but when it visits x_2 it discovers that the path to the cheese is blocked, then the next time x_1 is visited, the appropriate choice of action (going right) can be made,

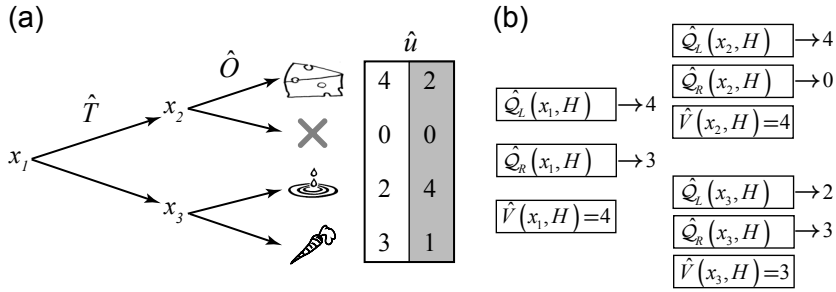

Figure 3.3 Complete maze task: (a) Goal-directed control uses a forward model of the transitions $\hat{T}$, the outcomes $\hat{O}$ and the utilities $\hat{u}$ to predict the values. (b) Cached control uses $\hat{Q}$ (and $\hat{V}$) values to quantify the worth of each action (and state) without modeling the underlying cause of these values. Figure adapted from Niv, Joel, and Dayan (2006).

since the new, impoverished consequence of going left can be calculated. This is a hallmark of the sort of sophisticated, cognitive control whose investigation was pioneered by Tolman (1948).

As mentioned, there is evidence for the involvement of dorsolateral prefrontal cortex and dorsomedial striatum in goal-directed control (Balleine 2005), perhaps because of the demands on working memory posed by forward search in the tree. The representation of $\hat{u}$, however, is less clear.

Habitual Control

Figure 3.3b is the extension of Figure 3.2b to the case of the full maze and shows the complete set of cached $\hat{V}$ and $\hat{Q}$ values for the case of hunger. These are actually the optimal values in the sense that they relate to the optimal choice of action at each state in the maze. As mentioned above, these values can be learned directly from experience. However, the cached values lack the flexibility of the forward model-based values, in particular not being based on knowledge, or estimates, of the actual outcomes. Therefore, behavior at x_3 may not change immediately with a change in motivational state from hunger to thirst, and the behavior at x_1 may not change immediately when the subject learns about the impossibility of turning left at x_2. These inflexibilities are the psychological hallmarks of automatic, habitual control (Sanfey et al. 2006).

One key feature for learning cached values in the case of trajectories is that the value of a state, or a prediction of this value, becomes a surrogate for the utilities that will arise for outcomes delivered downstream of that state. In the maze, for instance, assuming that the subject already knows to turn left there, state x_2 is worth $V(x_2, H) = 4$ in the case of hunger. This makes state x_2 a conditioned reinforcer (see Mackintosh 1983). Discovering this on turning left at x_1 can help reinforce or stamp in that choice at x_1. In the case of rewards, it is a learning rule that arises from this that has been validated in the activity

of dopamine neurons, the concentration of dopamine at its striatal targets, and even the fMRI BOLD signal in humans undergoing simple conditioning tasks for primary or secondary reward (Montague et al. 1996; Schultz 1998; O'Doherty 2004). The combination of learning predictions and using those predictions to learn an appropriate policy is sometimes called the actor–critic learning rule (Sutton and Barto 1998).

These and other studies suggest that habitual control depends strongly on dorsolateral regions of the striatum (Balleine 2005). This region is subject to dopaminergic (and serotonergic) neuromodulation, and there is evidence from pharmacological studies that these neuromodulators influence the learning of appropriate actions.

Episodic Control

Both goal-directed and habitual controllers can be seen as learning appropriate actions through the statistical process of measuring the correlations between actions and either outcomes or the utilities of those outcomes. An apparently more primitive scheme, called an episodic controller, would be simply to repeat wholesale any action or sequence of actions that has been successful in the past (Lengyel and Dayan 2007). This requires a structure that can store the successful actions, putatively a job for an episodic memory, plus a way of coupling ultimate success to storage or consolidation within this memory. Although different interpretations are possible, there is indeed evidence in rats and humans that the hippocampus, a critical structure for episodic memory, whose storage appears to be under rich neuromodulatory control, may indeed have a key role to play (Poldrack and Packard 2003).

Pavlovian Control

The three controllers described thus far encompass ways of making, or at least learning to make, essentially arbitrary choices in light of the outcomes to which they lead. They are sometimes called instrumental controllers, since they favor actions that are instrumental in achieving rewards or getting to safety. This sort of choice may, however, just be a thin layer of icing on the overall cake of control (Breland and Breland 1961). Inbuilt, instinctive mechanisms, which have been programmed by evolution, specify vast swathes of appropriate actions in the direct face of particular appetitive or aversive outcomes (Mackintosh 1983). This is particularly apparent in the response to clear and present dangers; such defensive and aggressive responses even appear to enjoy a topographic organization along an axis of the dorsal periacqueductal gray (Keay and Bandler 2001). It is also apparent in the appetitive (licking) and aversive (gapes) responses to the direct provision of primary rewards that Berridge and Robinson (1998) used to measure the extent of liking, as well as perhaps in

consummatory responses, such as the drinking behavior of a pigeon faced with a water dispenser or its eating behavior faced with grain in a hopper.

More critically, however, there is also a range of inbuilt responses to *predictions* of future rewards or punishments, which may be considered to come from a Pavlovian controller. These responses are emitted even though they may be immaterial (or indeed, as will be seen below, antagonistic) to the delivery of the outcomes. Very crudely, one component of these responses is outcome-independent, including approach and engagement to predictors associated with positive utilities and withdrawal and disengagement from predictors associated with negative utilities. For instance, pigeons will approach or peck a key whose illumination is temporally predictive of the delivery of food or water at another part of the experimental chamber; rats will move away from a stimulus that predicts the delivery of a shock. A second component is outcome-dependent; that is, the detailed topography of the keypeck is dependent on whether the pigeon expects food or water.

In view of these findings, it is likely that there are multiple Pavlovian controllers, with outcome-dependent responses arising from the predictions of something akin to the forward model, and outcome-independent reactions arising from something more like cached values. Indeed, exactly the same anatomical and pharmacological mechanisms seem to play similar roles, and it may be that Pavlovian values arise from the systems discussed above. However, the different classes of Pavlovian responses seem not to be separated as readily using motivational manipulations, and the distinctions between the different controllers is less well understood.

Pavlovian control is important since most tasks that are designed to test instrumental controllers also create Pavlovian contingencies (Mackintosh 1983). In the maze of Equation 1, for instance, how do we know if hungry animals run from x_1 to x_2 because they have had the choice of going left stamped in by the utility available (an instrumental explanation) or because they have learned to predict the appetitive characteristic of $\hat{V}(x_2) > 0$, and to emit an inbuilt approach response to stimuli or states associated with positive utility (a Pavlovian one)? In fact, a *reductio ad absurdum* exists: since Pavlovian responses have likely been designed over evolution to be appropriate to the general environmental niche occupied by the subjects, experimenters can arrange it such that the Pavlovian response is instrumentally inappropriate. A famous case of this is negative automaintenance: Consider the autoshaping experiment mentioned above, in which pigeons come to peck keys simply because they are illuminated a short while before food or water is presented. If an additional, instrumental, contingency is imposed, such that no food or water will be provided on any trial on which the pigeon pecks the illuminated key, the pigeons continue to peck the key to some degree, despite the adverse effect of this pecking on their earnings. Many foibles of control and choice can be interpreted in terms of untoward cooperation and competition between Pavlovian and instrumental control (Dayan et al. 2006).

Interactions among the Systems

It is natural to ask what benefits could accrue from the existence of multiple, different decision-making systems and, indeed, to question how the new problem for choice they pose collectively might be solved; that is, choice between the choosers rather than between different possible actions.

Instrumental Competition

The first three decision makers at least share the common notional goal of making choices that will maximize predicted utilities. They base their choices on distinct value systems, which are differentially accurate in given circumstances. Daw et al. (2005) suggested that goal-directed and habitual controllers lie at two ends of a rough spectrum, trading off two sources of uncertainty: ignorance and computational noise. This trade-off is not unique to animals, but rather affects all decision makers. Daw et al. (2005) suggested that uncertainty should also be the principle governing the choice between the decision makers.

The goal-directed controller makes best use of information from the world, propagating findings at one state (e.g., a newfound inability to turn left at x_2) immediately to affect choice at other states (enabling the subject to choose to turn right at x_1). By contrast, information propagates between the in a more cumbersome manner over learning, and thus incorrect or inefficient behaviors (at states like x_1) may persist for a number of trials. In fact, one possible use for the reactivation of activity during sleep or quiet wakefulness is the consolidatory process of propagating information to eliminate any inaccuracy. In general, the goal-directed controller can be expected to be more accurate early in learning and after environmental change. Its use of $\hat{u}$ also makes it more reliable in the face of change to the motivational state.

Nevertheless, the goal-directed controller only gains this flexibility at the expense of computationally intractable and, at least in animals, noisy, computations. The task of explicitly searching the tree of choices is tough, since the number of branches and leaves of a tree typically grows very quickly with its depth. Keeping accurate track of the intermediate states and values is impossible for deep trees, and the resulting inaccuracies are a source of noise and uncertainty that removes therefore some of the benefits of goal-directed control. Note that if computational noise can be reduced through the application of (cognitive) effort or processing time, then forms of effort- or time-accuracy trade-offs may ensue (e.g., Payne et al. 1988).

Computational intractability has inspired the integration of goal-directed and habitual control in artificial systems. For instance, game-playing programs, such as the chess player Deep Blue, typically perform explicit search, but without reaching the leaves of the tree (where one or another player will be known to have won). Instead, they use an evaluation function to measure the worth of the positions on the board discovered through searching the tree.

This evaluation function is exactly analogous to the $V(x, m)$ function discussed above. As before, the ideal evaluator would report something akin to the probability of winning, starting from position x. Indeed, some of the earliest ideas in reinforcement learning stem from the realization that the tree search defines a set of consistency conditions for the evaluations, and that inconsistencies can be used to learn or improve evaluators (Sutton and Barto 1988).

Daw et al. (2005) suggest that the goal-directed and habitual systems should keep track of their own respective uncertainties. Then, as is standard in Bayesian decision theoretic contexts, the controllers' estimates should be integrated in a way that depends on their uncertainties. Calculating the uncertainties actually poses severe computational challenges of its own, and it is likely that approximations will be necessary. Daw et al. (2005) demonstrate that a very simple version of uncertainty-based competition sufficed to account for the data on the transition from goal-directed to habitual control that is routinely observed over the course of behavioral learning. This transfer is signaled by the different motivational sensitivities of the two controllers, which, in turn, comes from their dependence on different value systems. It occurs because after substantial experience, the uncertainty in the habitual controller, which stems from its inefficient use of information, is reduced below the level of the computational noise in the goal-directed system. Daw et al. (2005) interpret various psychological results by noting that the extent to which uncertainty favors one or another system depends on the depth of the tree (shallow trees may permanently favor goal-directed control) as well as the (approximate) models of uncertainty that each controller possesses.

The episodic controller can also fit into this general scheme for uncertainty-based utility and competition. Repeating a previously successful action is likely to be optimal in the face of substantial uncertainty and costly exploration. This suggests that the episodic controller should be most beneficial at the very outset of learning, even before the goal-directed controller. Experimental data on this are presently rather sparse (Poldrack and Packard 2003).

Pavlovian Competition

The Pavlovian controller is somewhat harder to locate within this apparently normative interaction between the instrumental controllers. It is difficult, for example, to see the logic behind the finding in negative automaintenance that pigeons will *learn* to emit actions (pecking the illuminated key because of its temporal association with the delivery of food) that are explicitly at odds with their instrumental ambitions (i.e., denying them food). More strikingly, learning may be based on exactly the same forward-modeling and cached value systems that underlie the instrumental decision makers.

One way to think about this is as an illusion of choice. Some perceptual illusions can be seen as arising when a particular observation is constructed in a way that violates statistical expectations or priors based either on

evolutionarily specified characteristics of the natural sensory environment, or on characteristics that have been acquired over a lifetime of experience. The systems engaged in perceptual inference must weigh the evidence associated with a particular scene against the overwhelming evidence embodied by these priors. Illusions can result when the latter wins. Similarly, it is indeed generally appropriate to approach and engage with predictors of positive utility, and it is mainly in artificial experimental circumstances that this has negative consequences. Breland and Breland (1961) present an extensive and entertaining description of various other cases that can be seen in similar terms.

The task for us becomes one of understanding the net effects of the interaction between Pavlovian and instrumental control. Approach and engagement responses triggered by positive values could lead to the same sort of boost for a nearly immediately available rewarding outcome that underlies the impulsivity inherent in hyperbolic discounting (for a recent example, see McClure et al. 2004). Concomitantly, it could underlie indirectly the whole range of commitment behaviors that have been considered by proponents of this sort of discounting (e.g., Ainslie 2001).

Withdrawal responses caused by predictions of aversive outcomes can be equally important and, at least in experimentally contrived conditions, equally detrimental. However, perhaps not unreasonably, the range of Pavlovian responses to threat appears to be much more sophisticated than that to reward, with a need to engage flight (definitive withdrawal) or fight (definitive approach) under conditions that are only fairly subtly different. Nevertheless, it has been suggested (De Martino et al. 2006) that the framing effect, in which presenting equivalent choices as losses rather than gains makes them apparently less attractive, could well arise from Pavlovian withdrawal. Similarly, one can argue that Pavlovian responses might bias the sort of deliberative evaluation performed using a forward model, by inhibiting the exploration of paths that might have negative consequences and boosting paths that might have positive consequences. Various biases like these are known to exist, and indeed are systematically disrupted in affective diseases such as depression.

The last interaction is called Pavlovian-to-instrumental transfer, or PIT. This is the phenomenon whereby a subject engaging in an instrumental action to get an outcome will act more quickly or vigorously upon presentation of a Pavlovian predictor of that outcome, or indeed any other appetitive outcome that is relevant to the subject's current motivational state. The boost from predictors of the same outcome (called specific PIT) is greater than that of a different, though motivationally relevant, outcome (general PIT). The obverse phenomenon, in which a predictor of an aversive outcome will suppress ongoing instrumental behavior for rewards, is actually the standard way that the strengths of Pavlovian predictors of aversive outcomes are measured. One view of general PIT is that the Pavlovian cues might affect the prediction of the long-run average value ρ and, as discussed above, thereby change the optimal behavioral vigor (Niv, Joel, and Dayan 2006). A mechanism akin to this was

interpreted as playing a key role in controlling the vigor with which habits are executed. This was argued as being necessary since the general motivational insensitivity of habits limits other ways of controlling habitual vigor.

Open Issues

There are many open issues about the nature of the individual systems and their interactions. In particular, the nature of the predicted utility $\hat{u}_o(m)$ in the forward-model value, and especially its dependence on motivational state m and relationship to the native "liking" value $u_o(m)$, are distressingly unclear. Equally, the coupling between forward-model mechanisms and neuromodulators is somewhat mysterious—a fact that is particularly challenging for the models designed to optimize the average rate of utility provision. Subjects have also been assumed to know what actions are possible, although this in itself might require learning that could be particularly taxing.

I would like to end by pointing to various issues that relate to the broader themes of this Ernst Strüngmann Forum.

Generalization and Value

I have stressed issues that relate to generalization: the way that information learned in one motivational state generalizes to another one, and generalization of facts learned in one part of an environment (e.g., the futility of going left at x_2) to other parts (the choice of going right at x_1). The issue of generalization in learning functions such as $\hat{Q}_c(x, m)$ is a large topic itself, raising many questions about things such as the representations of the states x and m that determine these values.

Foremost to generalization is prior expectations. Subjects can have prior expectations about many facets of the world, which then translate into prior expectations about all the unknown aspects of the problem (particularly T and O). Take, for instance, the phenomenon of learned helplessness, which is an animal and human model of depression (e.g., Maier 1984). In learned helplessness, two subjects are yoked to receive exactly the same shocks. One subject (the "master") can terminate the shock by its actions, whereas the other (the "yoke") cannot; it just experiences whatever length of shock the master experiences. In generalization tests to other tasks (or other environments), the master subjects perform like controls; however, the yoked subjects show signs that resemble depression, notably an unwillingness to attempt to choose actions to improve their lot. One interpretation (Huys and Dayan, pers. comm.) is that the yoked subjects generalize to the test, the statistical structure, which indicates that actions cannot be expected to have reliable or beneficial effects. Given this expectation, it is not adaptive to search for or to attempt appropriate actions, which is exactly the behavioral observation. Thus predictions about

likely values and value structures can exert a strong influence on the whole interaction between a subject and its environment.

Conscious and Unconscious Control

A major theme at this Forum was the relationship between conscious and unconscious decision making. Without wishing to pass judgment on consciousness, we might think that the former has more to do with deliberative processing; the latter with automatic control. If we consider forward-model schemes to be deliberative and habitual ones automatic, then within the instrumental systems described herein, there appears to be a seamless integration with a single intended target. The Pavlovian controller can disturb this normativity, but it may also be based on deliberative and automatic evaluation mechanisms, each of which might disturb its matched instrumental controller. Our view is thus perhaps not completely consistent with some two-systems accounts, for instance, those that attribute a shorter timescale of discounting to the automatic rather than the deliberative system (McClure et al. 2004).

Deliberation, however, could clearly be much more complicated than the sort of rigid search that we have considered through a known tree. There are many choices inherent in the search process itself, and these are presumably the products of exactly the same sorts of mechanisms and rules that have been developed for externally directed actions. Understanding how they all work together as well as the way that biases in different systems affect the overall choice, are key areas for future exploration. To belabor one example, I have argued that amorphous and poorly understood tasks likely favor the episodic controller. However, since it is explicitly not designed on the principles of statistical averaging, which underlie the forward model and the habit system, this controller will most likely have a difficult time integrating statistical information correctly. This could lead to large biases in choice, given only moderate numbers of examples.

Conclusions

I have sketched an account of the influences of different systems involved in the assessment of value on different systems involved in making choices. Value systems, and therefore the behavioral systems they support, vary according to the nature and origin of their sensitivity to the motivational state of a subject, its knowledge about its environment, as well as their intrinsic timescales. Decision-making systems vary in related ways, bringing different information to bear on largely similar concerns. Instrumental controllers work together in a harmonious manner, whereas it is significantly more challenging to account for and understand Pavlovian control.

Acknowledgments

I am very grateful to my many collaborators in the work described here, notably Bernard Balleine, Nathaniel Daw, Yael Niv, John O'Doherty, Máté Lengyel, Ben Seymour, Quentin Huys, Ray Dolan, and Read Montague. I specially thank Nathaniel Daw, Yael Niv, and Ben Seymour for detailed comments on an earlier draft. I also thank Christoph Engel, Wolf Singer, and Julia Lupp for organizing the workshop, and to them, the members of Groups 1 and 2, Jon Cohen, and Eric Johnson for helpful comments. Funding was from the Gatsby Charitable Foundation.

References

Ainslie, G. 2001. Breakdown of Will. Cambridge: Cambridge Univ. Press.

Balleine, B. W. 2005. Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiol. Behav.* **86**:717–730.

Balleine, B. W., and S. Killcross. 2006. Parallel incentive processing: An integrated view of amygdala function. *Trends Neurosci.* **29**:272–279.

Berridge, K. C., and T. E. Robinson. 1998. What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Res. Rev*. **28**:309–369.

Breland, K., and M. Breland. 1961. The misbehavior of organisms. *Am. Psychol*. **16**:681–684.

Damasio, A. R., D. Tranel, and H. Damasio. 1991. Somatic markers and the guidance of behavior: Theory and preliminary testing. In: Frontal Lobe Function and Dysfunction, ed. H. S. Levin, H. M. Eisenberg, and A. L. Benton, pp. 217–229. New York: Oxford Univ. Press.

Daw, N. D., S. Kakade, and P. Dayan. 2002. Opponent interactions between serotonin and dopamine. *Neural Networks* **15**:603–616.

Daw, N. D., Y. Niv, and P. Dayan. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neurosci*. **8**:1704–1711.

Dayan, P., and B. W. Balleine. 2002. Reward, motivation and reinforcement learning. *Neuron* **36**:285–298.

Dayan, P., Y. Niv, B. J. Seymour, and N. D. Daw. 2006. The misbehavior of value and the discipline of the will. *Neural Networks* **19**:1153–1160.

De Martino, B., D. Kumaran, B. Seymour, and R. J. Dolan. 2006. Frames, biases, and rational decision-making in the human brain. *Science* **313**:684–687.

Dickinson, A. 1994. Instrumental conditioning. In: Animal Learning and Cognition, ed. N. Mackintosh, pp. 45–79. San Diego: Academic Press.

Dickinson, A., and B. Balleine. 2001. The role of learning in motivation. In: Stevens' Handbook of Experimental Psychology, vol. 3: Learning, Motivation and Emotion, 3rd ed., ed. C.R. Gallistel, pp. 497–533. New York: Wiley.

Kahneman, D., P. Wakker, and R. Sarin. 1997. Back to Bentham? Explorations of experienced utility. *Qtly. J. Econ.* **112**:375–406.

Keay, K. A., and R. Bandler. 2001. Parallel circuits mediating distinct emotional coping reactions to different types of stress. *Neurosci. Biobehav. Rev*. **25**:669–678.

Konorski, J. 1948. Conditioned Reflexes and Neuron Organization. Cambridge: Cambridge Univ. Press.

Konorski, J. 1967. Integrative Activity of the Brain: An Interdisciplinary Approach. Chicago: Univ. of Chicago Press.

Lengyel, M., and P. Dayan. 2007. Hippocampal contributions to control: The third way. Neural Information Processing Systems conference.

Loewenstein, G. 1996. Out of control: Visceral influences on behavior. *Org. Behav. Hum. Dec. Proc*. **65**:272–292.

Luce, R. D. 1959. Individual Choice Behavior: A Theoretical Analysis. New York: Wiley.

Mackintosh, N. J. 1983. Conditioning and Associative Learning. Oxford: Oxford Univ. Press.

Maier, S. F. 1984. Learned helplessness and animal models of depression. *Prog. Neuropsychopharm. Biol. Psych.* **8**:435–446.

McClure, S. M., D. I. Laibson, G. Loewenstein, and J. D. Cohen. 2004. Separate neural systems value immediate and delayed monetary rewards. *Science* **304**:503–507.

Montague, P. R., P. Dayan, and T. J. Sejnowski. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci*. **16**:1936–1947.

Niv, Y., N. D. Daw, D. Joel, and P. Dayan. 2006. Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharm.* doi:10.1007/s00213-006-0502-4.

Niv, Y., D. Joel, and P. Dayan. 2006. A normative perspective on motivation. *Trends Cogn. Sci.* **8**:375–381.

O'Doherty, J. P. 2004. Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Curr. Opin. Neurobiol*. **14**:769–776.

Payne, J. W., J. R. Bettman, and E. J. Johnson. 1988. Adaptive strategy selection in decision making. *J. Exp. Psychol.: Learn. Mem. Cogn.* **14**:534–552.

Poldrack, R. A., and M. G. Packard. 2003. Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia* **41**:245–251.

Sanfey, A. G., G. Loewenstein, S. M. McClure, and J. D. Cohen. 2006. Neuroeconomics: Cross-currents in research on decision-making. *Trends Cogn. Sci.* **10**:108–116.

Schultz, W. 1998. Predictive reward signal of dopamine neurons. *J. Neurophys*. **80**:1–27.

Sutton, R. S., and A. G. Barto. 1998. Reinforcement Learning. Cambridge, MA: MIT Press.

Tolman, E. C. 1948. Cognitive maps in rats and men. *Psychol. Rev*. **55**:189–208.

Watkins, C. J. C. H. 1989. Learning from delayed rewards. Ph.D. thesis, Univ. of Cambridge.