Available online at www.sciencedirect.com

ScienceDirect

Neural Networks 19 (2006) 1153–1160

Neural Networks

www.elsevier.com/locate/neunet

2006 Special Issue

The misbehavior of value and the discipline of the will

Peter Dayan[a,*], Yael Niv[a,b], Ben Seymour[c], Nathaniel D. Daw[a]

[a] *Gatsby Computational Neuroscience Unit, UCL, 17 Queen Square, London WC1N 3AR, United Kingdom*
[b] *ICNC, Hebrew University, PO Box 1255, Jerusalem 91904, Israel*
[c] *Wellcome Department of Imaging Neuroscience, UCL, 12 Queen Square, London WC1N 3BG, United Kingdom*

Received 31 October 2005; accepted 30 March 2006

Abstract

Most reinforcement learning models of animal conditioning operate under the convenient, though fictive, assumption that Pavlovian conditioning concerns prediction learning whereas instrumental conditioning concerns action learning. However, it is only through Pavlovian *responses* that Pavlovian prediction learning is evident, and these responses can act against the instrumental interests of the subjects. This can be seen in both experimental and natural circumstances. In this paper we study the consequences of importing this competition into a reinforcement learning context, and demonstrate the resulting effects in an omission schedule and a maze navigation task. The misbehavior created by Pavlovian values can be quite debilitating; we discuss how it may be disciplined.

Keywords: Classical conditioning; Instrumental conditioning; Autoshaping; Negative automaintenance; Reinforcement learning; Will

1. Introduction

Theories of animal learning rest on a fundamental distinction between two classes of procedure: Pavlovian and instrumental conditioning (see Mackintosh (1983)). Crudely, the difference concerns contingency. In a Pavlovian (or classical) procedure, an animal learns that a stimulus (such as the ringing of a bell) *predicts* a biologically significant outcome (such as the delivery of a piece of meat) which is made to happen regardless of the animal's actions. The characteristic behavioral responses (*e.g.* salivation) that result are taken to reflect directly the animal's learned expectations. In instrumental (operant) conditioning, however, the delivery of the outcome is made to be *contingent* on appropriate actions (*e.g.* leverpresses) being taken by the animal. Ambiguity between Pavlovian and instrumental influences arises in that many behaviors, such as locomotion, can evidently occur under either Pavlovian or instrumental control. In fact, virtually all conditioning situations involve both sorts of circumstance; and the two varieties of learning are thought to interact with one another in a number of ways.

Here we investigate one such interaction — direct competition for behavioral output. This sort of competition has hitherto eluded the reinforcement learning (RL Sutton and Barto (1998)) theories that nevertheless have wide application in modeling substantial issues in both classical and instrumental conditioning (e.g. Dayan and Balleine (2002), Doya (1999), Hikosaka et al. (1999), Houk, Adams, and Barto (1995), Montague, Dayan, and Sejnowski (1996), O'Doherty (2004), Schultz (1998), Suri and Schultz (1998) and Voorn, Vanderschuren, Groenewegen, Robbins, and Pennartz (2004).

One high point in the debate about the relative importance of instrumental and classical effects in controlling behavior was the development of a Pavlovian procedure called *autoshaping* (Brown & Jenkins, 1968). This originally involved the observation that when the delivery of food reward is accompanied by the timely illumination of a pecking key, pigeons come to approach and peck the key. Critically, this pecking occurs even though (as a Pavlovian procedure) the food is delivered regardless of whether or not the key is pecked. In fact, this procedure leads more swiftly to reliable key pecking than the instrumental equivalent of only rewarding the birds with food on trials on which they peck. Classical conditioning

[*] Corresponding address: University College London, Gatsby Computational Neuroscience Unit, Alexandra House 17 Queen Square, London WC1N 3AR, United Kingdom. Tel.: +44 0 20 7679 1175; fax: +44 0 20 7679 1173.

E-mail addresses: dayan@gatsby.ucl.ac.uk (P. Dayan), yaelniv@alice.nc.huji.ac.il (Y. Niv), bseymour@fil.ion.ucl.ac.uk (B. Seymour), daw@gatsby.ucl.ac.uk (N.D. Daw).

0893-6080/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2006.03.002

ideas such as autoshaping actually underlie many schemes for shaping particular, apparently instrumental, animal behaviors.

By contrast, a procedure called negative automaintenance (Williams & Williams, 1969) uses an omission schedule (Sheffield, 1965) to pit classical and instrumental conditioning against each other. In the version of this adapted from autoshaping, the pigeons are *denied* food on any trial in which they peck the lit key. In this case, the birds *still* peck the key (albeit to a reduced degree), thereby getting less food than they might. This persistence in pecking despite the instrumental contingency between withholding pecking and food shows that Pavlovian responding is formally independent from instrumental responding, since the Pavlovian peck is never reinforced. However, it is disturbing for standard instrumental conditioning notions, which typically do not place restrictions on the range of behaviors that can be controlled by reward contingencies. Further, as Dayan and Balleine (2002) pointed out, but did not fix, it has particular force against the formal instantiation of instrumental conditioning in terms of RL. RL accounts neither for the fact that a particular action (pecking) accompanies the mere prediction of food, nor for the fact that this action choice can be better (or perhaps worse) than the instrumentally appropriate choice (in this case, of not pecking).

Such an anomaly is merely the tip of a rococo iceberg. In a famous paper entitled *The misbehavior of organisms*, Breland and Breland (1961) described a variety of more exotic failures of conditioning procedures (see also Breland and Breland (1966)). For instance, animals that initially learn to deposit an object in a chute to obtain food, subsequently become hampered because of their inability to part with the food-predicting object. Equally, Hershberger (1986) showed that, in a 'looking glass' environment, chicks could not learn to run *away* from a source of food in order to get access to it. Many of these failures have the flavor of omission schedules, with an ecologically plausible action (the equivalent of approaching and pecking the lit key) interfering with the choices that would otherwise lead to desirable outcomes. Various of the behavioral anomalies arise *progressively*, with the instrumentally appropriate actions slowly being overwhelmed by Pavlovian ones.

Humans also exhibit behaviors that seem to violate their apparent goals. This has most frequently been studied in terms of a long-term plan (*e.g.* dieting) being bulldozed by a short-term opportunity (*e.g.* a cream bun). Indeed, this sort of intertemporal choice conflict lies at the heart of two popular theories. One theory suggests the conflict arises from hyperbolic discounting of the future (see Ainslie (1992, 2001), Laibson (1997), Loewenstein and Prelec (1992) and Myerson and Green (1995), which makes short term factors overwhelm a long term view. Another theory is that the behavioral anomalies arise from competition between deliberative and affective choice systems (Loewenstein & O'Donoghue, 2004; McClure, Laibson, Loewenstein, & Cohen, 2004), with the latter ignoring long-term goals in favor of immediate ones. However, data on interactions between deliberative and affective instrumental systems in animals are well explained (see Daw, Niv, and Dayan (2005)) by assuming the controllers actually share the *same* goals and differ only in terms of the information they bring to bear on achieving those goals. Therefore, here we propose that, instead of intertemporal conflicts being key, the anomalies may arise from interactions between Pavlovian control and instrumental control of either stripe. The appearance of intertemporal competition follows from the character of the Pavlovian responses, which seem myopic due to being physically directed toward accessible reinforcers and their predictors.

In this paper, we propose a formal RL account of the interaction between the apparently misbehaving Pavlovian responses (arising from classically conditioned value predictions) and instrumental action preferences. As mentioned, we have recently (Daw et al., 2005) studied competition in RL between multiple subsystems for instrumental control — a more reflective, 'goal-directed' controller and its 'habitual' counterpart; the present work extends this approach to the interactions between instrumental (for simplicity, here represented by a single habitual controller) and Pavlovian control. In Section 2, we show how our model gives rise to negative automaintenance in an omission schedule, and in Section 3, we explore the richer and more varied sorts of misbehavior that it produces in the context of a navigational task. Finally, Ainslie (1992, 2001), and following him Loewenstein and O'Donoghue (2004), consider the *will* as the faculty that allows (human) subjects to keep their long-term preferences from being derailed by short-term ones. We consider how the will may curb Pavlovian misbehavior.

2. Negative automaintenance

Consider first the simplest case of instrumental conditioning in which animals learn to execute action N (NOGO: *withholding* a key peck) which leads to reward $r = 1$ rather than action G (GO: *pecking* the key) which leads to reward $r = 0$. For convenience, we consider a $\mathcal{Q}$-learning scheme (Watkins, 1989) in which subjects acquire three quantities (all at trial t):

1. $v(t)$ the mean reward, learned as $v(0) = 0$, and
$$v(t+1) = v(t) + \eta(r(t) - v(t)) \tag{1}$$
where $r(t) \in \{0, 1\}$ is the reward delivered on trial t, and η is a learning rate. This is a simple instance of the Rescorla and Wagner (1972) rule, which is actually just the same as would be found in temporal difference learning (Sutton, 1988) in this context;
2. $q_{\mathsf{N}}(t)$ is the $\mathcal{Q}$ value of action N (*i.e.* the expected value of performing action N), updated as $q_{\mathsf{N}}(0) = 0$ and
$$q_{\mathsf{N}}(t+1) = q_{\mathsf{N}}(t) + \eta\,(r(t) - q_{\mathsf{N}}(t)) \tag{2}$$
only if N was chosen on trial t.
3. $q_{\mathsf{G}}(t)$ is the $\mathcal{Q}$ value of action G updated as $q_{\mathsf{G}}(0) = 0$ and
$$q_{\mathsf{G}}(t+1) = q_{\mathsf{G}}(t) + \eta\,(r(t) - q_{\mathsf{G}}(t)) \tag{3}$$
only if G was chosen on trial t.

The $\mathcal{Q}$ values as such determine the *instrumental* propensity to perform each of the actions, with action $a(t) \in \{\mathsf{N}, \mathsf{G}\}$ chosen according to

$$p(a(t) = \mathsf{N}) = \frac{e^{\mu(q_{\mathsf{N}}(t) - v(t))}}{e^{\mu(q_{\mathsf{N}}(t) - v(t))} + e^{\mu(q_{\mathsf{G}}(t) - v(t))}} \tag{4}$$
$$= \sigma\left(\mu(q_{\mathsf{N}}(t) - q_{\mathsf{G}}(t))\right) \tag{5}$$

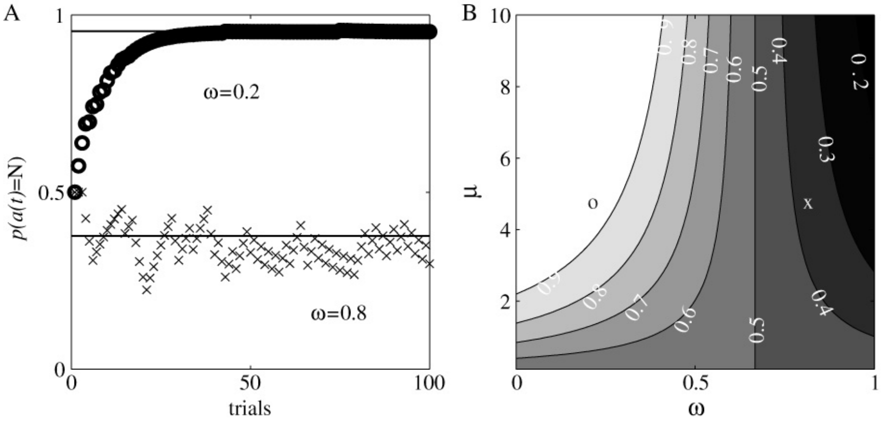

Fig. 1. Negative automaintenance. The figures show the consequences of a Pavlovian bias for an action (G) that leads to omission of reward. (A) The evolution of the probability $p(a(t) = \mathsf{N})$ over the course of learning for learning rate $\eta = 0.2$, competition parameter $\mu = 5$ and Pavlovian reliability $\omega = 0.2$ ('o') or $\omega = 0.8$ ('x'). The greater the weight accorded to the Pavlovian action, the stronger the omission effect. The solid horizontal lines show the theoretical equilibrium probabilities for the two values of ω. (B) Contour plot of the equilibrium values of the probability $p(a = \mathsf{N}) = v$ across μ and ω. The two values from (A) are marked by their respective symbols. The diminution of instrumental performance engendered by the Pavlovian influence is apparent.

where $\sigma(\cdot)$ is the standard logistic sigmoid function and μ is a parameter governing the steepness of the preference for the higher-valued action. This softmax function (or Luce choice rule, 1959) is just one of a number of possibilities for action competition — but is widely adopted in RL contexts. If $a(t) = \mathsf{N}$ is always followed by $r(t) = 1$ and $a(t) = \mathsf{G}$ by $r(t) = 0$, then $q_\mathsf{N}(t) \to 1$; $q_\mathsf{G}(t) \to 0$, action N is ultimately chosen a fraction $p_\mathsf{N} = \sigma(\mu)$ of the time and $v(t) \to p_\mathsf{N} = \sigma(\mu)$. For $\mu > 3$, $p_\mathsf{N} > 0.95$ and so the correct choice is dominant.

Note that the action probabilities in Eq. (5) do not depend on the state value $v(t)$, since this same quantity is subtracted from each Q-value in Eq. (4). The terms $q_\mathsf{N}(t) - v(t)$ and $q_\mathsf{G}(t) - v(t)$ in this equation are the *advantages* (Baird, 1993; Dayan & Balleine, 2002) of actions N and G. Advantages are closely associated with actor-critic control, and have also been directly used to model the neural basis of instrumental conditioning (O'Doherty et al., 2004).

We model the omission aspects of the schedule by suggesting that one of the actions is labelled as being the Pavlovian action. The advantage of this action is then augmented by an amount that depends on the Pavlovian *value* of the state $v(t)$, thus potentially distorting the competition with the instrumentally favored N. If the innate Pavlovian action were N (withholding), then the learning of the Pavlovian contingency (*i.e.* the predictive value of the state) would *speed* the course of instrumental learning. Indeed, such synergy between Pavlovian and instrumental goals is exactly why autoshaping can hasten instrumental acquisition. However, in the model of the negative automaintenance case, the Pavlovian action is G (pecking, as in typical appetitive approach behavior), which leads to competition between the Pavlovian and instrumental propensities. From an RL viewpoint, the choice between N and G is purely arbitrary, and is shaped by the external reward contingencies; however, these actions are not symmetric from a psychological (Pavlovian) viewpoint, and this leads exactly to the omission issue.

We need to formalize the competition between Pavlovian and instrumental actions. There are various ways to do this — one simple version is to treat the Pavlovian impetus towards G to be exactly the value $v(t)$ (and the Pavlovian impetus to N to be 0), and to consider weighting this Pavlovian factor with the instrumental advantages $q_\mathsf{N}(t) - v(t)$ and $q_\mathsf{G}(t) - v(t)$ for the two actions. More exactly, we weight the Pavlovian impetus by ω, and the instrumental advantages by $(1-\omega)$ where $0 \le \omega \le 1$ acts as if it is the assumed competitive *reliability* of the Pavlovian values (Dayan, Kakade, & Montague, 2000). We discuss the aetiology of ω more fully later; first we consider the consequences of choosing different values for it.

In sum, the propensities to perform each action are changed to:

$$\begin{aligned} &\mathsf{N}: q_\mathsf{N}(t) - v(t) \Rightarrow (1-\omega)(q_\mathsf{N}(t) - v(t)) \\ &\mathsf{G}: q_\mathsf{G}(t) - v(t) \Rightarrow (1-\omega)(q_\mathsf{G}(t) - v(t)) + \omega v(t) \end{aligned} \tag{6}$$

and action choice is performed as in Eq. (4), giving

$$p(a(t) = \mathsf{N}) = \sigma(\mu((1-\omega)(q_\mathsf{N}(t) - q_\mathsf{G}(t)) - \omega v(t))). \tag{7}$$

Fig. 1A shows the results of simulating this model, with $\eta = 0.1$, $\mu = 5$ and Pavlovian reliabilities $\omega = 0.2$ ('o') and $\omega = 0.8$ ('x'). The results show $p(a(t) = \mathsf{N})$ on a trial-by-trial basis. They start at $p(a(t) = \mathsf{N}) = 0.5$ (as the propensities to both actions are zero), and converge to final values, near to 1 for $\omega = 0.2$ and near to 0.4 for $\omega = 0.8$ (thin solid lines). This shows the basic effect of Pavlovian contingencies on performance in the omission schedule — the propensity to perform the appropriate action is greatly diminished by the Pavlovian bias for the 'wrong' action.

If behavior stabilizes (which depends on the learning rate), then $v(t) \to p(a(t) = \mathsf{N})$. We can therefore look for equilibrium points of Eq. (7) as a function of μ and ω. Fig. 1B shows these equilibrium values. Negative automaintenance bites more strongly as ω increases (*i.e.* $p(\mathsf{N}) = v$ gets smaller), an effect that is enhanced by increasing action competition. In this formulation, for $\omega = \frac{2}{3}$, it turns out that $v = \frac{1}{2}$, independent of μ. This defines the central vertical contour of that plot. The symbols 'o' and 'x' mark the locations explored through the simulations in Fig. 1A.

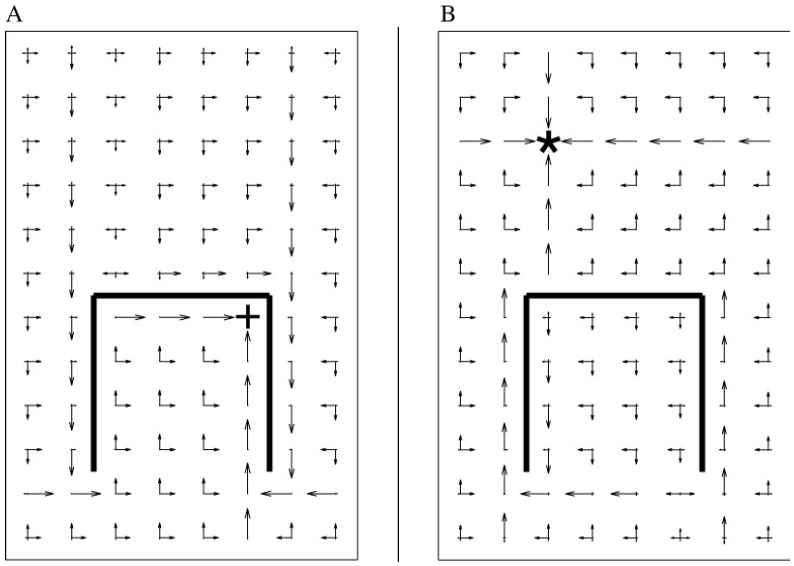

Fig. 2. Pavlovian and instrumental conditioning in a standard 8 × 12 maze with a fence. Both plots illustrate near-optimal stochastic policies taking a subject from any place in the maze to one of two different goals, with the aim of maximizing γ^T, where $\gamma = 0.9$ is a discount factor, and T is the number of steps. The policy is determined by the softmax competition of the advantages defined by these values, with $\mu = 50$ (since the values are very close to each other). Arrow length is proportional to probability.

In sum, as a didactic example, we have shown one way to incorporate paradoxical behavior on omission schedules within RL, by manipulating the advantages of the actions. The parameterization includes the weight ω, which determines the relative importance of the instrumental and classical contingencies. One might expect ω normally to be fairly high — since the generalizations that underlie classical conditioning's choice of actions have presumably been honed by evolution, and should thereby be accorded a high reliability. Thus a suitable role for an overarching control process would be to reduce ω, and thereby enable more proficient collection of rewards in relevant circumstances. Alternatively, ω could itself be subject to inference (perhaps according to modeled uncertainties about the relative reliabilities of Pavlovian and instrumental actions; Daw et al. (2005)).

3. Detours

Omission schedules, and indeed the interestingly florid behaviors exhibited by Breland and Breland (1961)'s actors or Hershberger's (1986) chicks, concern relatively constrained sets of actions. However, classical contingencies may exert a rather more all-pervasive influence over other sorts of behavior, warping choice according to the proximity of relatively immediate goals and their near precursors. We would argue that some of the many apparent illogicalities of choices, such as those studied under the rubric of emotional contributions to irrationality (*e.g.* Loewenstein and O'Donoghue (2004)) arise from this source. We come back to this point in the discussion, focusing particularly on choice anomalies arising from hyperbolic discounting that have been much studied by Ainslie (1992, 2001), Laibson (1997) and Loewenstein and Prelec (1992). In this section, we consider another very simplified model of how Pavlovian choices can warp instrumental responding, using navigation in a maze. We discuss later the reasons why mazes may not be quite the optimal model; however, we use them since the ready interpretability of policies in mazes makes them excellent didactic tools.

Fig. 2A illustrates a simple 8 × 12 grid maze (often used as an example in previous RL papers Barto, Sutton, and Watkins (1990)) with an 'ᑎ'-shaped fence. The agent can move in the four cardinal directions, except that if it attempts to cross the fence or stray outside the maze, then it just stays where it is. There is a goal just inside the fence, shown by the '+' sign, for instance, some food. If we consider the food as having a value of 1 unit, then the arrows illustrate a near-optimal, probabilistic policy for minimizing the number of steps to the food (or rather maximising the temporally discounted return $\gamma^T \cdot 1$ obtained, where $\gamma = 0.9$ is a discount factor, and T is the number of steps to the goal). The policy indicates which direction the agent takes, with the lengths of the arrows proportional to their probabilities. Formally, the probabilities come from the softmax generalization of Eq. (4), with each action competing according to its exponentiated advantage, where the $\mathcal{Q}$-value defining the advantage of an action again comes from estimates of the expected values $\mathcal{E}[\gamma^T]$ based on starting by taking that action (Watkins, 1989)).

To capture the misbehavior created by Pavlovian urges, we consider an additional, Pavlovian secondary conditioning goal, say a light previously associated with food. The light is presented at a different location in the maze (shown by the asterisk in Fig. 2B). In the same way that the key elicits approach and pecking from the hapless pigeons, the light elicits approach. For convenience (see later discussion), we use the optimal approach policy shown in the figure (which is actually created in the same way as the optimal policy for the true goal), as the Pavlovian action propensities.

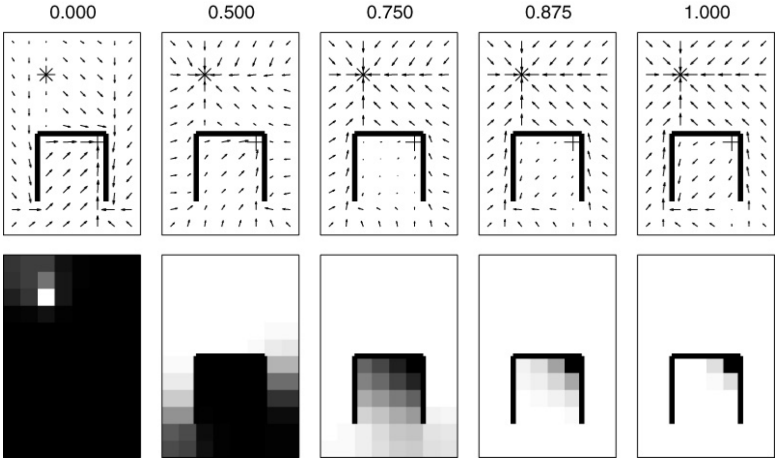

Fig. 3. Pavlovian warping. As a function of the weight ω accorded to the Pavlovian goal (the asterisk), top and bottom plots show the stochastic policies (for illustrative convenience, arrows are shown in the average direction at each point; in reality the agent stochastically chooses one single cardinal direction) and the probabilistic event horizons of the Pavlovian goal (on a grayscale of 0–1). For small ω, the instrumental goal dominates; as ω gets larger, the Pavlovian goal exerts a greater influence. Here $\mu = 40$.

Consider the effect of turning on the light (the asterisk) when the agent is moving (under instrumental control) to the plus sign. This creates a Pavlovian imperative to approach the light (for convenience, we assume this can be seen throughout the maze), which then competes with the instrumental control, and therefore warps the trajectories. We model this competition using the parameter ω, just as in Eq. (7), except that four actions compete rather than two, and the Pavlovian influence is governed by the Pavlovian advantages for approach (generating the approach policy shown in Fig. 2B). We use $\mu = 40$, and vary the reliability parameter ω as a way of titrating the control exerted by Pavlovian actions.

The top row of Fig. 3 shows how the (net) instrumental policy to the '+' is warped by the illumination of the Pavlovian goal as a function of parameter ω. For large ω, the whole policy turns around, and moves the agent towards the Pavlovian goal rather than the instrumental one. For intermediate ω, the effect of the light on the path taken by the agent depends on its starting point within the grid. One way to quantify the warping created by the Pavlovian goal is the extent to which trajectories to the '+' get incorrectly captured by the '$*$'. For a given pair of goals, we can calculate the probability (averaging over the stochastic choice of actions) that this will happen starting from any location in the maze. This results in a form of probabilistic 'event horizon' around the Pavlovian goal which consists of the starting locations from which the agent is likely to visit it first. The bottom row of Fig. 3 shows these event horizons as a function of ω. For large ω, only the Pavlovian goal has any force, and it seduces the agent from almost anywhere in the maze. Of course, one might expect that, having approached and engaged with the light, the agent will then set a course for the instrumental goal.

4. Discussion

In this paper, we have considered a simple, policy-blending interaction between Pavlovian and instrumental actions, using ideas from reinforcement learning to provide a formal framework. Pavlovian actions (such as approach to cues predicting rewards), which are presumably stamped in by their evolutionary appropriateness, can sometimes interfere negatively with instrumental goals, leading to poor control. This is starkly evident in omission schedules, which are designed to emphasize this competition. However, the navigational example in Section 3 suggests a more pervasive class of problems in which similar issues are important. We now consider more general aspects of Pavlovian-instrumental interactions, some areas in which our model is incomplete, and finally relate our study to recent work (Daw et al., 2005) on the computational structure of instrumental conditioning, and thereby to competing explanations of the sub-optimalities that we have considered.

4.1. Pavlovian–instrumental interactions

First, Pavlovian actions can certainly be synergistic with instrumental actions rather than antagonistic (a fact that is important in the debate about the independence of Pavlovian and instrumental conditioning, see Mackintosh (1983)). Indeed, in many experiments, Pavlovian actions are actively solicited, for instance, to encourage subjects to engage with manipulanda. Even in the maze task, if the light were placed near the instrumental goal, then approach to the light could benefit acquisition of the food. In cases such as these forms of shaping, the dynamics of the interaction between Pavlovian and instrumental actions are carefully managed; in some subjects and procedures one can induce instabilities or even oscillations (Williams & Williams, 1969) by injudicious or unfortunate competition. We did not simulate either the positive interactions or the dynamics here, or indeed the motivational import of Pavlovian predictors evident in paradigms such as Pavlovian–instrumental transfer or conditioned suppression (see Dickinson and Balleine (2001)).

A synergistic interaction that is more critical for RL is *conditioned reinforcement* (see Mackintosh (1974)). In this, a stimulus that has a Pavlovian association with reward becomes a target for instrumental action. For instance, animals can learn (instrumentally) to press a lever to illuminate a light that has historically been paired with food, even absent any food delivery during the instrumental training. Conditioned reinforcement is the basic mechanism of action learning for long-term goals in models such as the actor-critic (Barto, Sutton, & Anderson, 1983). Conditioned reinforcement suggests that in a task like the maze, the problem is not so much that the two systems have different goals, in the sense of attaching value to different world states or objects, but, rather, that they direct different actions toward common goals, with the instrumental system seeking to illuminate the light, and the Pavlovian system to approach it (when illuminated). This is an important distinction between our interpretation of such tasks and the view of self-control theorists such as Loewenstein and O'Donoghue (2004).

Pavlovian–instrumental interactions may also occur in the aversive domain, although it has perhaps been less easy to discern domain (Mowrer, 1947). One difficulty is that, whereas appetitive unconditioned stimuli almost always engage approach, aversive outcomes cause either withdrawal or freezing or, in extreme circumstances, even approach (*e.g.* fighting). In these circumstances, there is not a unitary Pavlovian response that can be used to construct an omission schedule. At the time that there was active examination of the possibility that all Pavlovian responding was really instrumental, or *vice versa,* there were some attempts to use omission-like schedules in aversive cases (*e.g.* Bolles, Stokes, and Younger (1966), Coleman (1975) and Kamin (1956). There is also, for instance, a report that squirrel monkeys punished for biting on a restraining leash tend to increase their biting (Morse, Mead, & Kelleher, 1967). However, there appears not to be a very wide range of aversive studies directly pitting Pavlovian against instrumental choices.

4.2. Model lacunæ

To make the illustration concrete, we had to specify a rather large number of factors about which there is presently little data and also work in a rather simplified regime. However, general aspects of the competition extend beyond these particular choices. In particular, mazes and spatial navigation tasks potentially involve very different learning mechanisms from more arbitrary instrumental tasks (*e.g.* Foster, Morris, and Dayan (2000) and O'Keefe and Nadel (1978). Here, we interpret the maze as a relatively rich example of a sequential decision problem, showing off the temporally extended consequences of competition between Pavlovian and instrumental systems. We also did not study learning in the maze; but rather assumed prior knowledge of Pavlovian and instrumental contingencies. The main issues for learning would be the same as those illustrated by the negative automaintenance example.

Further, we selected Pavlovian actions (to the light) according to their advantages, which are themselves rather instrumental. However, the idea is to blend two sets of action preferences (*i.e.* policies), and the advantages are really just a convenient way of specifying the more obvious geographical controller of moving directly towards the light. The latter would be similar to the advantages, except for the effect of the fence. The impact of fences and the like were studied extensively by Köhler (1925); Tolman (1948) and their followers. For instance, dogs (though apparently not chimpanzees) faced with food on the other side of a boomerang-shaped fence will run around the fence and eat the food if they are released from the leash far enough from the food, but when released close to the fence they run towards the food and thus fail to reach it (Köhler, 1925). This indeed suggests a powerful element of direct approach coming from the Pavlovian system that does not even accord with the structure of the maze (in this case, by being insensitive to the fence). This is certainly an important area for further experimental investigation (Foster, 2000).

Also, the use of a grid maze implies that there are many actions that are equivalent for one of the two particular goals, making it possible in some cases to choose actions that satisfy both. However, it is apparent from Fig. 3 that the effects of the two goals extend to actions that require opposite directions. Equally, the relative magnitudes of Pavlovian and instrumental advantages are rather arbitrary — here, the light and the goal had the same nominal reinforcement magnitude. Motivational manipulations of either or both goals would be particularly interesting, because of a debate (Daw et al., 2005; Dayan & Balleine, 2002; Dickinson & Balleine, 2001; Holland, 2004) as to the extent to which Pavlovian values (and thus presumably advantages) directly reflect motivational states, without the need for learning.

4.3. Progressive anomalies, habitization and self-control

In many of the cases reported by Breland and Breland (1961), behavioral anomalies arise progressively — with subjects learning to be proficient at instrumental tasks before innately specified actions take over. This is tantalizingly reminiscent of the transfer of control of instrumental behavior from outcome-value sensitive, goal-based, control to outcome-value *in*sensitive, habitual control. Importing into RL a rather widespread view (Dickinson, 1985; Owen, 1997; Packard & Knowlton, 2002), we have recently suggested (Daw et al., 2005) that goal-directed action (associated with dorsolateral prefrontal cortex, dorsomedial thalamus, and various of their afferents and efferents) arises from search in an explicit model of the task and environment, whereas habitual action (associated with the amygdala and the dorsolateral striatum) arises from cached or stored versions of values, policies or advantages, with learning being determined by neuromodulators. In our theory, these controllers compete based on their own estimates of their reliability, with goal-directed control having an advantage early in training (because it uses samples more statistically efficiently) but later ceding it to habitual control (because of inaccuracies arising from sources such as computational complexity).

If goal-directed control indeed arises from tree-search, then the deleterious consequences of Pavlovian actions will be explicit, and they may therefore be easier to eliminate. The habitual system does not use an explicit model of the outcomes of actions, and so Pavlovian and instrumental advantages would be on a common footing, and would directly compete. Thus we may expect Pavlovian effects to be more parasitic on habitual than goal-directed control. This would explain why the Pavlovian effects arise with training, as goal directed gives way to habitual control. This parasitism could readily be tested, using behavioral neuroscience techniques (such as lesions of infralimbic and prelimbic cortices or the dorsal striatum; Balleine and Dickinson (1998), Coutureau and Killcross (2003) and Killcross and Coutureau (2003)) that are known to manipulate the relative contributions of goal-directed and habitual control to behavior.

This interaction between Pavlovian actions, and goal-directed and habitual instrumental behavior is a main difference between our view and that of Loewenstein and O'Donoghue (2004). They suggest that choices determined by *deliberation* compete with choices determined *affectively*, and interpret many self-control situations (such as when a subject must forgo proffered food to achieve some other ends) accordingly. Willpower can be expensively exerted by the deliberative system to overwhelm affective choices, but its costs have to be weighed against its benefits. In application to intertemporal choice, they suggest that the deliberative system has both long and short term goals, but the affective system only has the latter (further reinforced by immediate proximity to primal goals). In the particular case that the cost of *future* willpower is ignored, distal outcomes turn out to be discounted hyperbolically, suggesting an alternative explanation for at least some of the extensive behavioral phenomena discussed by Ainslie and others (*e.g.* Ainslie (1992, 2001), Laibson (1997) and Loewenstein and Prelec (1992).

Relating Loewenstein and O'Donoghue (2004)'s suggestion to ours, we suggest that the equivalent of their affective and deliberative systems are our Pavlovian and instrumental systems, respectively. However, Loewenstein & O'Donoghue's nomenclature elides the key subdivision (Daw et al., 2005) in the instrumental system between habitual and goal-directed control, only the latter of which might really be called deliberative, and both of which have different sorts of motivational or affective sensitivities. A somewhat deeper interpretational difference is that our Pavlovian system does not necessarily have different *goals* from our instrumental systems (in fact, here we have accorded it just the same goals), but rather that it has a somewhat rigid and impoverished set of actions (largely approach or withdrawal ultimately consummated in ingestion or removal) that it brings to bear on achieving those goals. It is these actions that often give it an apparently myopic character (though this need not always be so, as demonstrated by the long-range event horizons in the maze of Fig. 3).

As Dickinson and Balleine (2001) have extensively discussed, there are many subtleties to motivationally-sensitive control in Pavlovian and instrumental systems, all of which can affect the nature and outcome of the competition. In particular, Pavlovian approach, based on motivationally inappropriate goals (sustained by goal-independent values) can disrupt instrumental actions. In an example like the maze, this could have subjects interrupting their instrumentally determined actions (say toward a cognitive goal such as work) on account of being seduced by Pavlovian approach to a cream bun at a bakery on the route. It could also have their whole choice of route warped by the implicit imperative of getting close to the bakery itself.

Due to these competing myopic actions, our model also exhibits some of the inconsistent choices and short-termism that Loewenstein and O'Donoghue (2004), and also Ainslie (2001), identify in their models as being related to hyperbolic discounting. However, we do not suggest that Pavlovian actions are wholly responsible for all hyperbolic discounting — distinctly non-exponential discounting functions are widely found even in purely cognitive tasks (Cropper, Aydede, & Portney, 1991; Myerson & Green, 1995), where there are no apparent competing Pavlovian actions.

Finally, the degree to which Pavlovian responses are useful (favouring high values of the parameter ω) may largely be determined on an evolutionary timescale. However, to the extent that this parameter is modifiable, we would expect it to change as experience accumulates about the success of Pavlovian responses in a particular environment, indeed, much in the same way we have suggested (Daw et al., 2005) that competition between goal-directed and habitual instrumental responses rests on inferences about their relative accuracies. As there, a great deal will turn on the tradeoff between the computational simplicity of Pavlovian control and the danger of suboptimal misbehavior. Whatever detailed principles actually govern the competition, when simple Pavlovian (and in fact also habitual control) lose out to more effortful, goal-directed control, the result can be viewed as a sort of top-down quashing. This, in turn, parallels the conception of Ainslie (2001) and Loewenstein and O'Donoghue (2004) of 'the will' as the faculty providing discipline (and favouring hard mental work) in the face of ease and temptation.

Acknowledgements

The authors are supported by the Gatsby Charitable Foundation (PD, NDD, YN), the Royal Society (NDD) the EU BIBA project (PD, NDD), a Dan David Fellowship (YN) and a Wellcome Trust program grant to Prof Ray Dolan (BS). We are very grateful to Kenji Doya and the Okinawa Computational Neuroscience Course for most pleasantly facilitating our interaction, and to two anonymous reviewers for their helpful comments.

References

Ainslie, G. (1992). *Picoeconomics*. Cambridge, England: Cambridge University Press.

Ainslie, G. (2001). *Breakdown of will*. Cambridge, England: Cambridge University Press.

Baird, L. C. (1993). Advantage updating. Technical report WL-TR-93-1146. Wright-Patterson Air Force Base.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*, 407–419.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike elements that can solve difficult learning problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *13*, 834–846.

Barto, A. G., Sutton, R. S., & Watkins, C. J. C. H. (1990). Learning and sequential decision making. In M. Gabriel, & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 539–602). Cambridge, MA: MIT Press.

Bolles, R. C., Stokes, L. W., & Younger, M. S. (1966). Does CS termination reinforce avoidance behavior? *Journal of Comparative and Physiological Psychology*, *62*, 201–207.

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, *16*, 681–684.

Breland, K., & Breland, M. (1966). *Animal behavior*. London, England: Macmillan.

Brown, P. L., & Jenkins, H. M. (1968). Auto-shaping of the pigeon's key-peck. *Journal of the Experimental Analysis of Behavior*, *11*, 1–8.

Coleman, S. R. (1975). Consequences of response-contingent change in unconditioned stimulus intensity upon the rabbit (*Oryctolagus cuniculus*) nictitating membrane response. *Journal of Comparative and Physiological Psychology*, *88*, 591–595.

Coutureau, E., & Killcross, S. (2003). Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behavioral Brain Research*, *146*, 167–174.

Cropper, M., Aydede, S., & Portney, P. (1991). Discounting human lives. *American Journal of Agricultural Economics*, *73*, 1410–1415.

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

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

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3*, 1218–1223.

Dickinson, A. (1985). Actions and habits — the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London*, *B308*, 67–78.

Dickinson, A., & Balleine, B. (2001). The role of learning in motivation. In C. R. Gallistel (Ed.), *Steven's handbook of experimental psychology*: *Vol 3*. *Learning, motivation and emotion* (3rd ed.). New York, NY: Wiley.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex? *Neural Networks*, *12*, 961–974.

Foster, D. J. (2000). A computational inquiry into navigation, with particular reference to the hippocampus. Ph.D. thesis. Department of Neuroscience, University of Edinburgh, Scotland.

Foster, D. J., Morris, R. G. M., & Dayan, P. (2000). Models of hippocampally dependent navigation using the temporal difference learning rule. *Hippocampus*, *10*, 1–16.

Hershberger, W. A. (1986). An approach through the looking-glass. *Animal Learning and Behavior*, *14*, 443–451.

Hikosaka, O., Nakahara, H., Rand, M. K., Sakai, K., Lu, X., Nakamura, K., et al. (1999). Parallel neural networks for learning sequential procedures. *Trends in Neurosciences*, *22*, 464–471.

Holland, P. C. (2004). Relations between Pavlovian-instrumental transfer and reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, *30*, 104–117.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davies, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge, MA: MIT Press.

Kamin, L. J. (1956). The effects of termination of the CS and avoidance of the US on avoidance learning. *Journal of Comparative and Physiological Psychology*, *49*, 420–424.

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, *13*, 400–408.

Köhler, W. (1925). *The mentality of apes*. New York, NY: Harcourt, Brace.

Laibson, D. I. (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, *62*, 443–478.

Loewenstein, G., & O'Donoghue, T. (2004). Animal spirits: Affective and deliberative processes in economic behavior. http://ssrn.com/abstract=539843.

Loewenstein, G., & Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *Quarterly Journal of Economics*, *57*, 573–598.

Luce, R. D. (1959). *Individual choice behavior*. New York, NY: Wiley.

Mackintosh, N. J. (1974). *The psychology of animal learning*. London, England: Academic Press.

Mackintosh, N. J. (1983). *Conditioning and associative learning*. Oxford: Oxford University Press.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, *306*, 503–507.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.

Morse, W. H., Mead, R. N., & Kelleher, R. T. (1967). Modulation of elicited behavior by a fixed-interval electric shock presentation. *Science*, *157*, 215–217.

Mowrer, O. H. (1947). On the dual nature of learning: A reinterpretation of conditioning and problem solving. *Harvard Educational Review*, *17*, 102–148.

Myerson, J., & Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior*, *64*, 263–276.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology*, *14*, 769–776.

O'Doherty, J., Dayan, P., Schultz, J., Deischmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*, 452–454.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. London: Clarendon.

Owen, A. M. (1997). Cognitive planning in humans: Neuropsychological, neuroanatomical and neuropharmacological perspectives. *Progress in Neurobiology*, *53*, 431–450.

Packard, M. G., & Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience*, *25*, 563–593.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–69). New York: Appleton-Century-Crofts.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1–27.

Sheffield, F. D. (1965). Relation between classical and instrumental conditioning. In W. F. Prokasy (Ed.), *Classical conditioning* (pp. 302–322). New York, NY: Appelton-Century-Crofts.

Suri, R. E., & Schultz, W. (1998). Abstract Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research*, *121*, 350–354.

Sutton, R. S. (1988). Learning to predict by the methods of temporal difference. *Machine Learning*, *3*, 9–44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*, 189–208.

Voorn, P., Vanderschuren, L. J., Groenewegen, H. J., Robbins, T. W., & Pennartz, C. M. (2004). Putting a spin on the dorsal-ventral divide of the striatum. *Trends in Neuroscience*, *27*, 468–474.

Watkins, C.J.C.H. (1989). Learning from delayed rewards. Ph.D. thesis. University of Cambridge, Cambridge, UK.

Williams, D. R., & Williams, H. (1969). Auto-maintenance in the pigeon: Sustained pecking despite contingent non-reinforcement. *Journal of the Experimental Analysis of Behavior*, *12*, 511–520.