Psychopharmacology
DOI 10.1007/s00213-006-0502-4

ORIGINAL INVESTIGATION

Tonic dopamine: opportunity costs and the control of response vigor

Yael Niv · Nathaniel D. Daw · Daphna Joel · Peter Dayan

Received: 27 April 2006 / Accepted: 28 June 2006

Abstract
Rationale Dopamine neurotransmission has long been known to exert a powerful influence over the vigor, strength, or rate of responding. However, there exists no clear understanding of the computational foundation for this effect; predominant accounts of dopamine's computational function focus on a role for phasic dopamine in controlling the discrete selection between different actions and have nothing to say about response vigor or indeed the free-operant tasks in which it is typically measured.
Objectives We seek to accommodate free-operant behavioral tasks within the realm of models of optimal control and thereby capture how dopaminergic and motivational manipulations affect response vigor.
Methods We construct an average reward reinforcement learning model in which subjects choose both which action to perform and also the latency with which to perform it. Optimal control balances the costs of acting quickly against the benefits of getting reward earlier and thereby chooses a best response latency.
Results In this framework, the long-run average rate of reward plays a key role as an opportunity cost and mediates motivational influences on rates and vigor of responding. We review evidence suggesting that the average reward rate is reported by tonic levels of dopamine putatively in the nucleus accumbens.
Conclusions Our extension of reinforcement learning models to free-operant tasks unites psychologically and computationally inspired ideas about the role of tonic dopamine in striatum, explaining from a normative point of view why higher levels of dopamine might be associated with more vigorous responding.

Keywords Dopamine · Motivation · Response rate · Energizing · Reinforcement learning · Free operant

Y. Niv (✉)
Interdisciplinary Center for Neural Computation,
The Hebrew University of Jerusalem,
Jerusalem 91904, Israel
e-mail: yaelniv@alice.nc.huji.ac.il

Y. Niv · N. D. Daw · P. Dayan
Gatsby Computational Neuroscience Unit,
University College London,
17 Queen Square,
WC1N 3AR London, UK

N. D. Daw
e-mail: daw@gatsby.ucl.ac.uk

P. Dayan
e-mail: dayan@gatsby.ucl.ac.uk

D. Joel
Department of Psychology, Tel Aviv University,
Tel Aviv 69978, Israel
e-mail: djoel@post.tau.ac.il

Introduction

Dopamine is perhaps the most intensively studied neuromodulator due to its critical involvement in normal behaviors, including learning and performance in appetitive conditioning tasks, and also in a variety of abnormal behaviors such as addiction, electrical self-stimulation, and numerous neurological and psychiatric disorders. Influenced particularly by the dramatic effects of pharma-

cological manipulations of dopamine neurotransmission on response rates, psychological theories of the neuromodulator's function have long focused on a putative role in modulating the vigor of behavior. These theories attribute the vigor effects to a variety of underlying psychological mechanisms, including incentive salience (Beninger 1983; Berridge and Robinson 1998; Ikemoto and Panksepp 1999), Pavlovian–instrumental interactions (Dickinson et al. 2000; Murschall and Hauber 2006), and effort–benefit tradeoffs (Salamone and Correa 2002). However, despite their psychological foundations, these theories do not, in general, offer a computational or normative understanding for why dopaminergic manipulations might exert such influence over response vigor.

A different influential line of empirical and theoretical work on the involvement of dopamine in appetitive conditioning tasks arose from electrophysiological recordings of midbrain dopamine neurons in awake, behaving monkeys. These recordings suggested that the phasic (bursting and pausing) spiking activity of dopamine cells reports to the striatum a specific "prediction error" signal (Ljungberg et al. 1992; Schultz et al. 1993; Schultz 1998; Waelti et al. 2001). Computational models showed that this signal can be used efficiently both for learning to predict rewards and for learning to choose actions so as to maximize reward intake (Sutton and Barto 1990; Friston et al. 1994; Barto 1995; Montague et al. 1996; Schultz et al. 1997).

However, these computational theories suffer from three deficiencies that prevent them from providing a comprehensive picture of the role of dopamine in conditioned responding: first, because they only treat the choice between discrete actions, they say nothing about the strength or vigor of responding. These models are therefore not capable of addressing free-operant behavior. Barring the interesting exception of McClure et al. (2003), which we discuss later, they also say nothing about the most obvious behavioral effect of pharmacological manipulations of dopamine, namely, their profound impact on response vigor.

Second, the computational theories generally assume that dopamine influences behavior only indirectly by controlling learning (e.g., Wickens 1990; Wickens and Kötter 1995). Although some behavioral effects of low-dose dopaminergic drug manipulations indeed emerge gradually, as if by learning (Wise 2004), more immediate effects are seen with higher drug doses (or medial forebrain bundle stimulation; Gallistel et al. 1974), and it seems implausible that dopaminergic drug effects are, in general, wholly mediated by learning (Ikemoto and Panksepp 1999).

Finally, whereas the unit recording data and associated computational theories are only concerned with the phasic release of dopamine, the tonic level of dopamine constitutes a potentially distinct and carefully controlled channel of neurotransmission (Grace 1991; Floresco et al. 2003; Bergstrom and Garris 2003; Goto and Grace 2005) for which a key role in enabling (Schultz 1998) or energizing (Weiner and Joel 2002) behavior has been suggested. Indeed, dopamine alterations affect a wide range of behaviors, many of which do not seem to be accompanied by phasic activity in dopamine cells. Furthermore, dopamine agonists can reverse many behavioral effects of dopamine loss, although they probably do not fully restore dopamine phasic transmission (Le Moal and Simon 1991; Schultz 1998). More directly, dopamine agonists or artificial increases in dopamine level (e.g., using amphetamine) have been shown to invigorate a range of behaviors (Lyon and Robbins 1975; Evenden and Robbins 1983; Taylor and Robbins 1984, 1986; Ljungberg and Enquist 1987).

Here we suggest that these three lacunæ are interrelated and can be jointly addressed. We do so by proposing a normative account of response vigor which extends the conventional computational view from discrete-choice, discrete-trial tasks to a more general continuous-time setting. We assume that animals choose the latency, time, or vigor with which they perform an action as well as which action actually to perform. We show that optimal decision making in the new framework has exactly the characteristics expected from psychological studies of the motivational sensitivity of response rates, including accommodating such apparent anomalies as hungry animals behaving more avidly even when performing actions (such as lever-pressing for water) that are not directed toward food gathering (Niv et al. 2005a, 2006).

The new theoretical model utilizes one new signal, namely, the average rate of reward, which we designate $\overline{R}$. The average rate of reward exerts significant influence over overall response propensities largely by acting as an opportunity cost, which quantifies the cost of sloth. That is, if the average rate of reward is high, every second in which a reward is not delivered is costly, and therefore, it is worth subjects' while performing actions more speedily even if the energetic costs of doing so are greater. The converse is true if the average rate of reward is low.

In the following, we first detail the extension of the standard model of learned action choice to the case of free-operant tasks, which brings about the need for this signal, and describe the results regarding its effects on response rates. We then argue on computational, psychopharmacological, and neural grounds that this average reward rate may be reported by tonic levels of dopamine, putatively in the nucleus accumbens, and show how it can account, without mediation through learning, for a wealth of reported effects of dopamine manipulations on response vigor in a variety of tasks. Finally, we consider how tonic and phasic dopamine signaling may interact in controlling behavior.

Methods: modeling response choice in free-operant tasks

Reinforcement learning (RL) is a computational framework for understanding how animals can predict future rewards and punishments, and choose actions that optimize those affective consequences (Sutton and Barto 1998). Not only does RL have a sound mathematical basis in the engineering theory of dynamic programming (Bertsekas and Tsitsiklis 1996), it also has long had a very close relation with psychological accounts of behavioral learning (Sutton and Barto 1981). Furthermore, RL offers a formal treatment of the phasic activity of dopamine neurons in primate ventral tegmental area (VTA) and substantia nigra pars compacta during appetitive conditioning tasks (Montague et al. 1996; Schultz et al. 1997). Briefly, it seems that phasic dopamine projections to the nucleus accumbens (as well as to the amygdala and prefrontal areas) report a form of prediction error about future rewards that is used to learn predictions of those future rewards. A similar phasic dopamine signal conveyed from the substantia nigra to the dorsal striatum seems to be involved in the adaptation of habitual actions to maximize future rewards (Packard and Knowlton 2002; Yin et al. 2004; Faure et al. 2005; Daw et al. 2005).

Almost all existing applications of RL have been to discrete-trial tasks, in which the only choices that subjects make are between different punctate actions (such as pressing either the left or the right lever in an operant chamber or running either left or right in a maze). This is clearly inadequate as a model of free-operant tasks, in which the key dependent variable has to do with when or at what rate an animal performs an action, in the light of different schedules of reinforcement and the animal's motivational (e.g., deprivational) state (Domjan 2003). Indeed, behavioral results indicate a delicate interplay between the costs of behaving faster and the possible benefits in terms of obtaining more rewards. This interplay results, for instance, in slower response rates the higher the interval or ratio schedule (Herrnstein 1970; Barrett and Stanley 1980; Mazur 1983; Killeen 1995; Foster et al. 1997) and faster responding on ratio schedules compared with yoked interval schedules (Zuriff 1970; Catania et al. 1977; Dawson and Dickinson 1990). Existing RL models also fail to capture key issues in discrete trial tasks for which the (e.g., energetic) costs of actions are balanced against their appetitive benefits (Cousins et al. 1996; Salamone and Correa 2002).

Here we suggest an extension to the standard RL model to the case that, along with making a choice between different possible actions, subjects also choose the latency (interpreted as response strength or vigor) with which they perform it (Niv et al. 2005a). Formalizing this allows the new model to accommodate all the issues raised above. The model may seem rather abstract and removed from either the behavior or the neural substrate. However, most of the definitions are directly related to the specification of the behavioral task itself. Furthermore, our abstraction of the optimizing task for the subject in a free-operant setting directly extends and parallels the abstraction of discrete-choice tasks in standard RL that has previously led to an account of psychological and neural data (Friston et al. 1994; Houk et al. 1995; Montague et al. 1996; Schultz et al. 1997). The model's abstractions and dynamics are summarized in Fig. 1 and are described below. A more detailed computational description can be found in the "Appendix".

We start by considering a simple free-operant task in which a (simulated) rat is placed in an operant chamber containing one lever and a food magazine. Several actions are possible: lever pressing (LP), nose poking (NP), and an action we will call "Other", which includes the range of other things that rats do in such scenarios (e.g., grooming, sniffing, and rearing). Food pellets fall into the food magazine as a result of lever pressing according to a designated schedule of reinforcement such as a fixed or random ratio or interval schedule (Domjan 2003). For simplicity, we assume that the rat can hear the food pellet falling into the magazine and therefore knows when it is actually available to be harvested via a nose-poke action.

We significantly simplify the dynamics of the interaction between the rat and the task by considering punctate choices. That is, at decision points, the rat chooses an action (a=NP, LP, or "Other") and the latency τ with which to perform this chosen action. Time τ then passes with no other actions allowed (the critical simplification), after which the action is completed. Latency (or rather inverse latency) is intended to formalize vigor—to complete a lever press within a shorter time, the animal must work harder. Following the period τ, any rewards that are immediately available for the action are harvested, and pellets scheduled to fall into the magazine do so. This may lead to a change in the state of the environment as observed by the rat. The rat then chooses another action and latency pair (a,τ), and the process continues. To allow comparison with experimental results (Foster et al. 1997), we also assume that eating a reward pellet is itself time-consuming; thus, whenever a nose-poking action is chosen and a pellet is available in the magazine, a variable "eating time" with a mean of several seconds (Niv et al. 2005b), must pass before the rat can make its next (a,τ) choice.

To complete the formal specification of the task for the rat, we have to describe the costs of performing actions, the utilities associated with the rewards, and the goal for the rat in the sense of what we consider it to be optimizing. We assume that each chosen action incurs both a fixed per-unit cost and a latency-dependent vigor cost (Staddon 2001), the

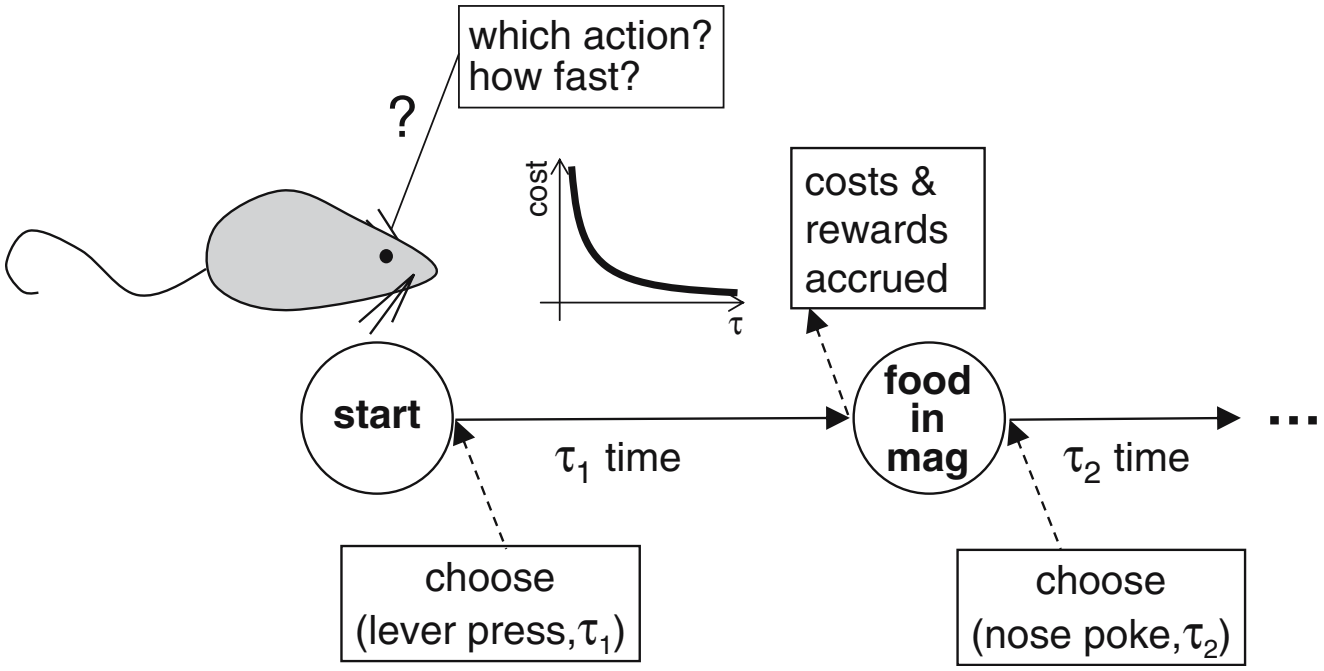

Fig. 1 Model dynamics. The simulated rat begins at the start state and selects both an action a to perform and the speed (i.e., the latency τ) with which to perform it. For instance, here, the rat selects the (action, latency) pair of (lever press, τ_1). As a simplification, we assume that the lever-press action is then executed (to the exclusion of all other actions) throughout the chosen latency τ_1. The action is completed at the end of this period, at which point all costs and available rewards are accrued and any state changes in the environment occur. In the ensuing state, the rat again selects an (a,τ) pair, and the process repeats

latter being proportional to the vigor of execution, that is, inversely proportional to the latency τ. Both costs can depend on the identity of the previous and the current action, for instance, so the cost of lever pressing after nose poking can take account of the unit and vigor costs of traveling from the magazine to the lever.[1] Together, these costs mean that lever pressing three times in 1 s is more costly than lever pressing three times in 3 s, and furthermore, lever pressing with a latency of 1 s after a previous lever press is less costly than lever pressing with the same latency after a nose poke. That some of the costs are vigor-related means that response rate selection is not trivial—responding too quickly is very costly, whereas responding too slowly means rewards will be obtained only after a considerable delay.

To model motivational manipulations such as hunger and satiety, we assume that the utility of the reward is dependent on the motivational (deprivational) state. That is, because we are modeling the delivery of pellets of food, we assume that their utility scales with hunger—food pellets will naturally be more valuable to a hungrier rat than to a sated one. Unfortunately, available data do not pin down the form of this mapping precisely, and we set it arbitrarily (for instance, a pellet might be worth three times as much to a hungry rat as to a sated rat). For simplicity, we also ignore small changes in reward utility potentially resulting from progressive satiation over the course of a session.

Last, but not least, we formalize the problem for the rats in terms of optimizing the long-run average rate of (net) utility per unit time. That is, the goal of the rat is to choose actions and latencies that maximize the accumulated rewards less the incurred costs, per unit time. This is a slightly different formulation from the more common exponentially discounted form of RL and is better suited for modeling many aspects of free-running, free-operant behavioral tasks which have a cyclic nature, and are not externally divided into trials and intertrial intervals (Daw 2003; Daw and Touretzky 2002; Daw et al. 2002). This average-reward RL formulation has also previously (Kacelnik 1997) been related to hyperbolic discounting (Ainslie 1975), which, in behavioral tests, typically bests exponential discounting.

In summary, we use RL techniques to determine the value of each pair (defined as the future expected rewards when taking this action with this latency compared to the average expected reward; Mahadevan 1996), at each state in the task. Once these values are known, the rat can achieve optimal performance simply by choosing, in every state, the (a,τ) pair with the highest value. The values can be

[1] Given that the actions we include in "Other" are typically performed in experimental scenarios despite not being rewarded by the experimenter, we assume these entail some "internal" reward, modeled simply as a negative unit cost.

computed using only local information (see "Appendix" for details). There are also online learning rules for acquiring the values based on experiencing many trials (Schwartz 1993; Mahadevan 1996), and indeed, those learning rules have been used as models of phasic and tonic dopamine and serotonin signaling (Daw 2003; Daw and Touretzky 2002; Daw et al. 2002). However, in this paper, we concentrate mostly on properties of the optimal solution, that is, how we expect well-trained animals to behave in the steady state, and leave the detailed time course of learning to future work.

Results: average reward rate and tonic dopamine

Figure 2 shows that optimal action selection under this scheme reproduces many characteristics of free-operant responding that are seen experimentally (Niv et al. 2005a). For instance, response rates are greater when rewards have higher values and are lower when the interval or ratio requirement of the reinforcement schedule is larger, according to the well-known hyperbolic relation (Catania and Reynolds 1968; see Fig. 2a,b). Furthermore, if two levers are available, each rewarding on a separate random-interval schedule, the model's behavior (Fig. 2d) is consistent with the ubiquitous experimental finding that the ratio of response rates on each lever matches the ratio of their reward rates (Herrnstein 1970; see Fig. 2c). Other characteristics such as faster responding on ratio schedules compared with yoked interval schedules (Zuriff 1970; Catania et al. 1977; Dawson and Dickinson 1990) are also reproduced by the model (Niv et al. 2005a).

In our model, the numerical value $\overline{R}$ of the average net reward per unit time plays a critical role in vigor selection,

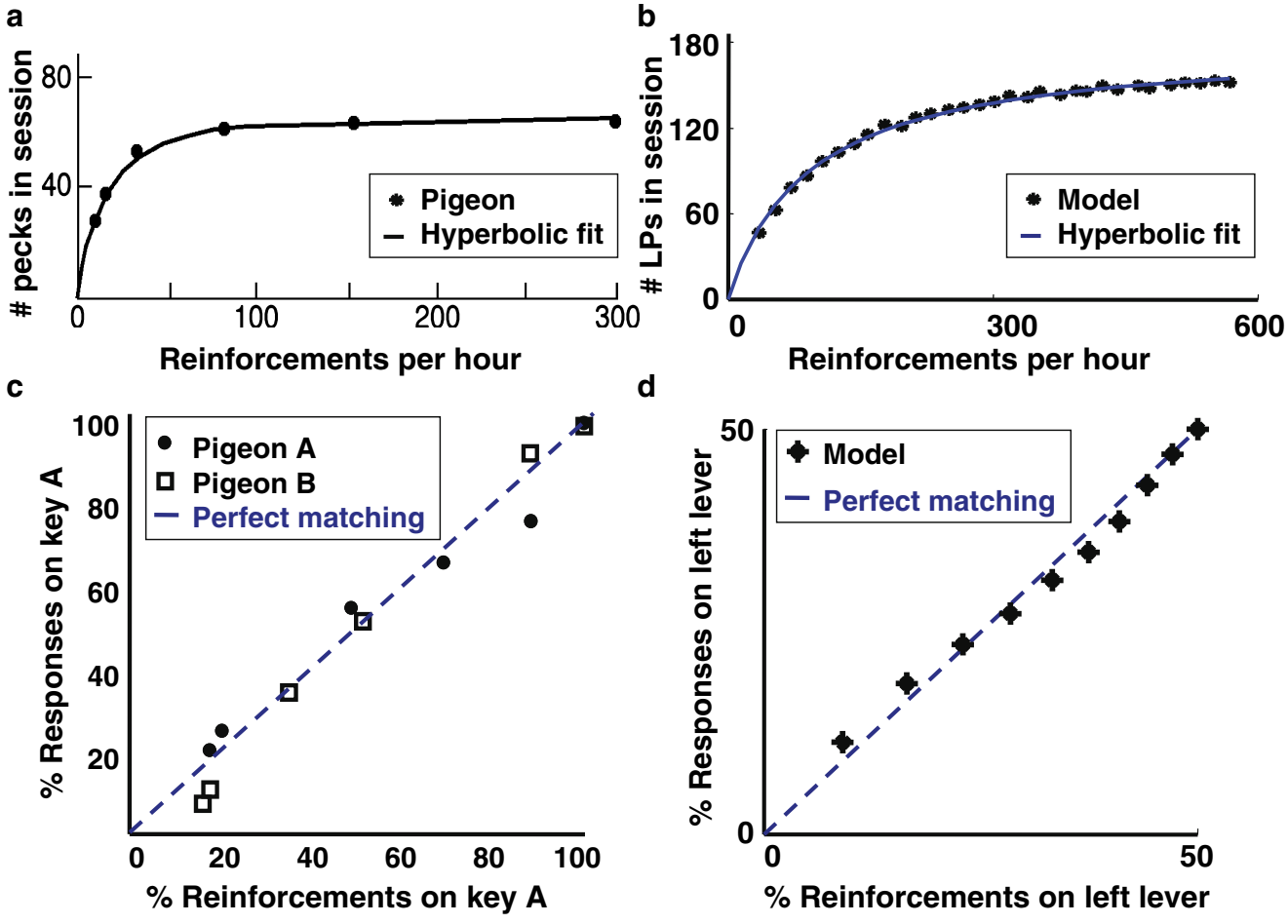

Fig. 2 Free operant behavior. **a** The relation between rate of responding (here, key pecking by a pigeon) and rate of reinforcement (the reciprocal of the interval between reinforcements) on a random-interval schedule (adapted from Herrnstein 1970; data originally from Catania and Reynolds 1968). **b** Model simulations capture the essence of the behavioral data. The relation between the total number of responses in a 30-min session (*circles*) and the rate of reinforcement is hyperbolic (*solid line*; hyperbolic curve fit). **c** Operant choice between two options, each reinforced on a separate random-interval schedule, follows the classic "Matching Law" (Herrnstein 1961) by which the proportion of responses (here, key pecking by pigeons) on one of the keys is equal to the proportion of rewards on that key (adapted from Herrnstein 1961). **d** Model simulations on a similar two-lever concurrent random-interval schedule also show matching behavior

acting as what might be seen as an opportunity cost. This is because when the rat chooses the latency with which to perform an action, it is occupied exclusively with that action for this entire duration. The cost of this commitment is $\tau \cdot \overline{R}$ because the rat is effectively forgoing this much reward on average by doing nothing other than the action it chose. Of course, if the chosen action is expected to yield more reward than this cost, it might still be worthwhile performing it. Thus, when choosing actions and latencies, the opportunity cost has to be weighed together with the cost of performing actions more quickly or vigorously and the benefit of getting the actual rewards sooner. In this way, the average reward introduces competition between different latencies of responding (Dragoi and Staddon 1999).

Analysis of the optimal solution in standard operant reinforcement schedules shows that the optimal latency of all actions is inversely proportional to the average reward rate ("Appendix"), a relation that turns out to be very revealing. It means that when the average reward rate is higher, optimal responding will be faster, and conversely, when the reward rate is lower, responding will be slower. This result can explain the well-known observation that hungrier rats are more "jumpy," performing all actions at a faster pace. The model shows that such counterintuitive, seemingly energy-wasting behavior is actually optimal: when hungry, the food pellets are subjectively more valuable, and thus, the overall expected reward rate in the experiment is higher. Thus, any action that is performed incurs a higher opportunity cost per unit time, implying a shorter optimal latency for all actions. Even when choosing an action such as grooming ("Other"), which does not lead to the coveted food reward, the optimal latency for this action will be shorter than that for a sated rat—intuitively, the rat should perform it quickly to resume lever pressing for the valuable food as soon as possible. Indeed, when hungry, the model rats not only choose to perform more food-obtaining actions, and at shorter latencies (resulting in a higher rate of lever pressing), but in addition, when they choose to perform "Other", they do this with a shorter latency as well. In this way, the model explains a type of "generalized drive" effect of motivation (Bolles 1967).

We therefore posit, on computational grounds, a slowly changing, tonic average reward signal that should exert control over generalized response vigor. Given the link between tonic dopamine and energizing of behavior (Weiner and Joel 2002), we propose that tonic dopamine carries this average reward signal. Much experimentation shows that higher levels of striatal dopamine are first and foremost associated with enhanced responsivity (Jackson et al. 1975; Carr and White 1987; Williams et al. Submitted) even before any learning-related effects are seen (Ikemoto and Panksepp 1999). Conversely, striatal dopamine depletion or antagonism profoundly reduces rates of responding (Sokolowski and Salamone 1998; Aberman and Salamone 1999; Salamone et al. 2001; Correa et al. 2002; Mingote et al. 2005). These direct effects of dopaminergic manipulations are exactly what is seen in our model if the average rate of reward is elevated or reduced. Therefore, the model provides a computational and normative foundation for understanding these effects and a bridge to psychological theories concerning them. Related to this interpretation, Montague (2006) has suggested that the lack of voluntary initiated movement in later stages of Parkinson's disease might be a normative consequence of a reduced expectation of average reward.

If we do indeed identify the average reward rate with tonic dopamine, we can now explicitly model the cost–benefit tradeoff experiments pioneered by Salamone and his colleagues (Salamone and Correa 2002). In the free-operant variant of these, it has been shown that 6-hydroxydopamine lesions in the accumbens have minimal effects on responding on low fixed-rate (FR) schedules while severely reducing responding on high FR schedules (Aberman and Salamone 1999; Salamone et al. 2001; Mingote et al. 2005). Figure 3a shows a representative experimental result. Figure 3b shows the result of the simulation in our model, with dopamine depletion simulated by reducing the average reward rate $\overline{R}$ while leaving all other aspects of the model intact ("Appendix"). A similar pattern of results is seen, with dopamine-depleted rats lever pressing less than control rats, an effect that is more pronounced for higher ratio schedules. This arises because the optimal latencies for lever pressing are longer once tonic dopamine reports a lower expected average reward rate. Thus, fewer presses are performed throughout the 30-min time allotted.

Note, however, that the apparently small effect on lever press rates in lower ratio schedules actually results in the simulation from a greater proportion of the session time spent eating in these schedules (at a consumption speed that is unaffected by dopamine depletions; Sokolowski and Salamone 1998; Aberman and Salamone 1999; Salamone et al. 2001; Salamone and Correa 2002) and not from a smaller effect on lever-press latencies. In the model, tonic dopamine ($\overline{R}$) depletion causes longer lever-press latencies in all schedules. However, in a schedule such as FR1, in which the rat performs a few hundred lever presses and is rewarded with several hundred pellets, the majority of the session time is spent consuming rewards rather than lever pressing for them. By comparison, in the FR16 condition, the rat presses over a thousand times and only obtains several tens of pellets; thus, effects of the dopamine depletion treatment on lever pressing seem more prominent.

It is noteworthy that dopamine depletions in our model also result in less switching between different actions because of the slower responding. This arises because slower actions incur lower vigor costs, against which the

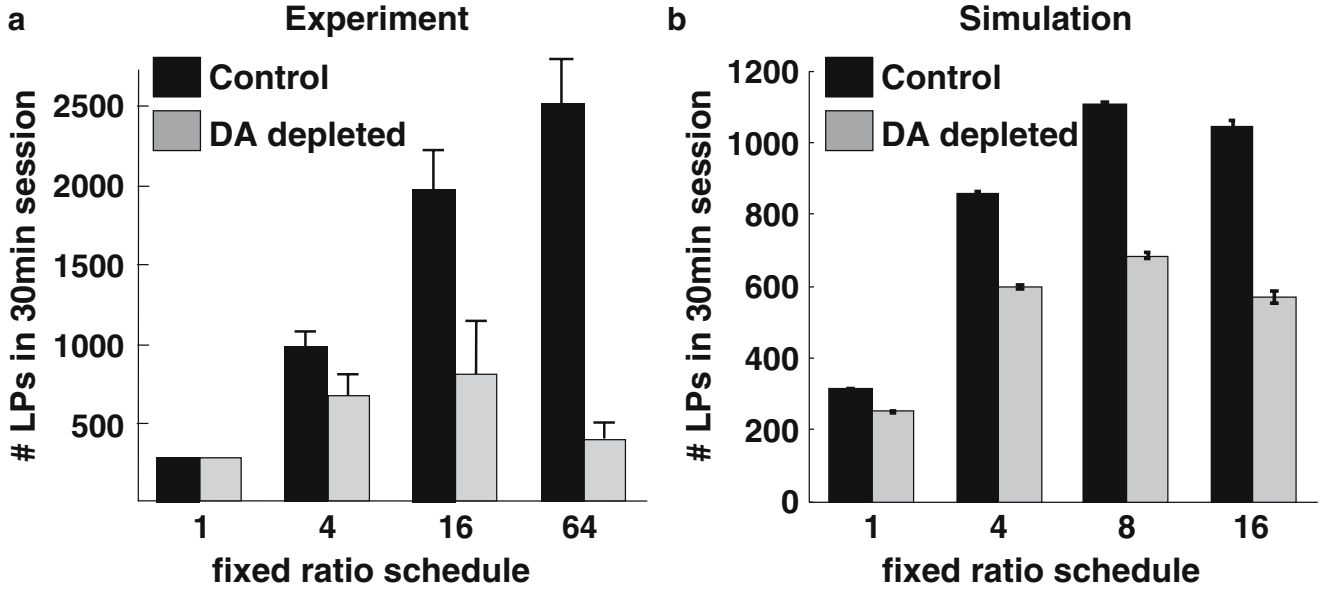

Fig. 3 Effects of dopamine depletion on fixed ratio responding. **a** Total number of lever presses per session, averaged over five 30-min sessions (*error bars*, SEM), by rats pressing for food on different fixed-ratio schedules. Rats with nucleus accumbens 6-hydroxydopamine lesions (*gray*) press significantly less than control rats (*black*), with the difference larger for higher ratio requirements (adapted from Aberman and Salamone 1999). **b** Simulation of dopamine depletion: overall lever press count over 30-min sessions (each bar averaging 15 sessions; *error bars*, SEM) for different fixed-ratio requirements. *Black* is the control condition, and *gray* is simulated dopamine depletion, attained by lowering the average reward rate $\overline{R}$ by 60%. The effects of the depletion appear more pronounced in higher schedules, but this actually results from an interaction with time spent feeding (see text)

impact of the vigor-independent switching costs loom comparatively larger. Action selection will therefore be biased toward repeating the currently chosen action, resulting in a free-operant form of perseveration. Indeed, lesion and drug studies have shown that dopamine loss results in perseveration, whereas increased dopamine promotes behavioral switching (e.g., Taghzouti et al. 1985; van den Bos et al. 1991), leading several researchers to suggest a role for dopamine in the control of switching (e.g., Robbins and Everitt 1982; Oades 1985; Weiner 1990; Le Moal and Simon 1991; Redgrave et al. 1999; Weiner and Joel 2002).

Discussion

We have presented a computational model of optimal action selection in free-operant tasks, incorporating the important, but often neglected, aspect of response vigor into an RL framework. Our model highlights the importance of the average rate of reward in determining optimal response rates and shows that higher reward rates should normatively be associated with faster responding and lower rates with slower responding. We suggest that the average reward rate is encoded by tonic levels of dopamine. This explains why this neuromodulator plays a critical role in determining the vigor of responding and provides a route by which dopamine could mediate the effects of motivation on vigor. This theory dovetails neatly with both computational theories which suggest that the phasic activity of dopamine neurons reports appetitive prediction errors, and psychological theories about dopamine's role in energizing responses.

Our framework puts in center stage the tradeoffs between costs and benefits that are explicit in the so-called cost–benefit T-maze tasks (Cousins et al. 1996), are implicit in free-operant tasks, and are manifest in day-to-day decisions about the vigor of actions. As these tradeoffs are typically continuous rather than binary, we suggest that rather than asking whether an animal is willing to put the effort into performing an action or not, we should ask how willing it is. In a recent model directed at somewhat similar data, McClure et al. (2003) made an explicit connection between response vigor and RL models of phasic, rather than tonic, dopamine function. Their theory effectively constructed a continuously varying response vigor (running speed) from a long series of binary decisions of whether to respond or to do nothing. This allowed them to incorporate effects of (phasic) dopamine on response vigor but did not license the sort of analysis of the tradeoff between response effort and benefit on which we have focused here.

Predictions

Our theory makes several readily testable predictions. First, we predict that lever-pressing latencies will be affected by

accumbens dopamine depletions of the sort employed by Salamone and colleagues, even in schedules requiring less effort per reward. This effect may not be straightforward to measure, however, because a molecular measure of response latency is needed rather than the molar measure of number of responses in a session. Indeed, a more detailed reaction time analysis in Mingote et al. (2005) points in this direction. One option would be to test effects of dopamine depletion during extinction to remove interactions with eating time. This would nicely separate immediate effects of changes in tonic dopamine levels from those of new learning due to a diminished phasic signal (see below), but albeit potentially at the expense of an interaction with extinction learning. Alternatively, higher-order schedules could be used to look at responding for conditioned stimuli, thereby eliminating the interference of rewards without inducing extinction.

We also predict similar effects of changes in motivational state. In particular, the higher the state of deprivation, the shorter the latency of all actions should be. Again, it would be important here to use a molecular measure of response latency to distinguish the effects of satiety on response rates from its effects on eating time (which, in this case, do appear to be significant; Aberman and Salamone 1999). Moreover, we predict that tonic levels of striatal dopamine will be higher in a deprived state than in a sated state (as also suggested by Weiner and Joel 2002), given that the animal has reason to expect a higher overall reward rate in its motivated state. Although difficult to measure directly, there is some supportive evidence for this (Wilson et al. 1995; Hernandez et al. 2006).

Immediate vs learned effects

Previous RL models have mostly concentrated on how phasic dopamine can affect behavioral preferences gradually and indirectly through a learning process. In contrast, we have modeled steady-state behavior in a well-learned task and focused on explaining how a change in tonic dopamine, caused either pharmacologically or by a change in deprivational state, can also affect behavior directly and immediately without requiring learning. The idea is that the system can take advantage of the fact that a higher average reward rate (arising, for instance, from a shift from satiety to hunger) will necessarily produce more vigorous optimal responding. It can then adjust response vigor directly on the basis of the tonic dopamine-reported average reward rate signal even before the new values of different (a,τ) pairs in the new situation have been learned. Importantly, such a mechanism provides some flexibility in rapidly adapting the overall level of behavior to changes in circumstance that are associated with changes in expected average reward rates.

Of course, the decision of how vigorously to respond is only one of the twin decisions underlying behavior in our framework. The decision as to which action to perform in a new motivational state is more difficult to adjust because it requires reestimating or relearning the values of different actions. In RL models of the sort we have considered, relearning involves additional training experience and utilizes the phasic dopamine signal. Thus, for instance, if a rat lever-pressing for food is shifted from hunger to thirst, the system will need new experience (and learning mediated by phasic dopamine) to direct its responding to an altogether different action to receive water. This complicated combination of direct motivational sensitivity (of vigor, through the tonic dopamine signal) and insensitivity (of choice, as a result of required learning) turns out to match well the results of experiments on a particular psychological category of "habitual" or "stimulus–response" behaviors (Dickinson 1985; Dickinson and Balleine 2002; Niv et al. 2006). Moreover, these are indeed associated with dopamine and the striatum (e.g., Yin et al. 2004; Faure et al. 2005).

We have not considered here the anatomically and psychologically distinct category of "goal-directed" behaviors (Dickinson and Balleine 1994), whose pattern of immediate and learned motivational sensitivity is rather different, and to which another class of RL models is more appropriate (Daw et al. 2005). Although our current model addresses habit-based instrumental control, optimizing the vigor of responding is as much an issue for goal-directed instrumental control and, indeed, for Pavlovian actions, and it is possible that average reward rates play a part in determining vigor for these as well.

We also have not treated here the learning of a new task but concentrated only on the steady-state situation. It is at this stage, in which responding is nearly optimal with respect to the reinforcement schedule and task, that we may analyze the optimal interrelation between reward rate and response vigor, and that these variables might stably be measured experimentally.[2] In contrast, learning is characterized by progressive (and likely complex) changes both in behavior and in the obtained average reward rate. Over the course of learning, the animal must continually estimate the average reward rate primarily from recently obtained rewards and costs. We predict that this estimate will control the dynamically changing tonic dopamine levels. In general, throughout acquisition, we can expect the experienced average reward rate to increase as the subject learns

[2] Realistically, even in a well-learned task, the average reward rate and response rates may not be perfectly stable. For instance, during a session, both would decline progressively as satiety reduces the utility of obtained rewards. However, this is negligible in most free-operant scenarios in which sessions are short or sparsely rewarded.

the contingencies and optimizes responding, obtaining more rewards while reducing action costs. Higher average reward rates reported by higher levels of tonic dopamine will further enhance response rates in a feedback cycle.

Tonic and phasic dopamine

Our model does not specify a mechanism by which tonic dopamine levels come to match the expected average reward rate. Various immediate and learned factors are likely to be important, and the complex pattern of interaction between tonic and phasic dopamine (e.g., Grace 1991; Moore et al. 1999; Floresco et al. 2003; Phillips et al. 2003; Goto and Grace 2005; Lodge and Grace 2005) appears to be key. One simple computational truth is that if the phasic responses of dopamine neurons indeed report a prediction error for future reward, their integration over time should, by definition, equal the average reward rate received. Indeed, phasic dopamine signals are discernable outside the synaptic cleft (Phillips and Wightman 2004; Roitman et al. 2004), so if tonic dopamine concentrations were solely determined by the slow accumulation of dopamine from phasic events, filtered by reuptake, they could directly measure throughout behavior the long-term cumulative average reward signal we posit. However, a view of the tonic signal as just the running average of the phasic signals is probably incomplete on both computational and physiological grounds. Computationally, we should expect the tonic average reward signal to be used predictively and not only reactively, which would require it to be somewhat decoupled from the actual obtained phasic reward signal. This would allow events such as changes in deprivation state to give rise to an immediate change in behavioral vigor based on previous learning of the relation between deprivation and expected average reward rate and before actually obtaining and averaging over rewards in the new motivational state.

Physiologically, there is indeed evidence that the two signals are somewhat decoupled, with phasic signals resulting from bursting activity and tonic dopamine levels determined mainly by the overall percentage of active (nonsilent) dopaminergic neurons and by presynaptic glutamatergic inputs (Chéramy et al. 1990; Chesselet 1990; Grace 1991; Floresco et al. 2003; Lodge and Grace 2006) (although these two modes of activity also interact in a facilitatory manner; Lodge and Grace 2005). Moreover, input structures to the VTA (e.g., the pedunculopontine nucleus and the ventral pallidum, respectively) appear to affect either bursting activity or population activity in dopamine neurons (Floresco et al. 2003; Goto and Grace 2005; Lodge and Grace 2006), providing another mechanism for independent modulation of phasic and tonic dopamine levels.

Because pharmacological manipulations of dopamine are likely to affect both tonic and phasic signaling, their effects on behavior can often be subtle to tease apart. This is illustrated, for instance, by considering our interpretation of responding in the cost–benefit T-maze task of Salamone and colleagues (Cousins et al. 1996; Denk et al. 2005). In this task, a rat can obtain four food pellets by choosing one arm of a T maze or one food pellet by choosing the other arm. However, the highly rewarding arm is partly blocked by a barrier which the rat must scale to reach the reward. Hungry rats typically choose the high-rewarding arm on the majority of trials. In contrast, after nucleus accumbens dopamine depletion, they prefer the low-rewarding arm, which demands less effort. We suggest that this reversal in discrete-action propensities is due to learned effects on choice preferences, mediated by the phasic dopamine signal. In particular, if the phasic signal is blunted by the drug, this would reduce the efficacy on learning of the four-pellet reward signal, making it, say, equivalent to the learning signal that would normally be seen on receiving only two food pellets. Of course, the reward signal for the low-rewarding arm would also be blunted, say, to the equivalent of half a pellet. In this case, although the three-pellet difference in reward before the lesion was sufficient to justify the extra cost of scaling the barrier, the 1.5-pellet difference after dopamine depletion might not, thereby altering the rats' choice toward the low-rewarding arm within a few trials of learning.

At the same time, we would expect reduced tonic dopamine to reduce the reported opportunity cost of slower responding without affecting the other aspects of the task. Thus, before new learning, rats should still be willing to climb the barrier for four food pellets; however, they need not hurry to do so. As a result of this separation of tonic and phasic effects, we predict that transiently, for instance, in the first postdepletion choice trial, dopamine-depleted rats should maintain their preference for the high-rewarding barrier arm, albeit acting distinctly more slothfully. Even as the phasic-induced learning effects accumulate and produce a shift in discrete action choice, the tonic effects should persistently promote slower responding. Indeed, Denk et al. (2005; see also Walton et al. 2006) show that dopamine depletion significantly lengthened the latencies to reach the reward when the barrier arm was chosen.

A final interaction between tonic and phasic aspects of dopamine is the finding that responding to cues predictive of higher reward is typically faster than responding to less valuable cues (e.g., Watanabe et al. 2001; Takikawa et al. 2002; Lauwereyns et al. 2002; Schoenbaum et al. 2003). Although vigor is generally associated with tonic rather than phasic dopamine in our model, unit recordings have shown a linear relation between reaction times and phasic dopaminergic responding (Satoh et al. 2003; see also Roitman et al.

2004). One possible explanation for these effects is that phasic signals transiently affect dopamine tone (Phillips and Wightman 2004; Roitman et al. 2004; Wise 2004), influencing vigor selection. Larger phasic prediction-error signals for stimuli previously associated with higher rewards (Fiorillo et al. 2003; Tobler et al. 2005) would then result in faster responding to these cues. Such effects of "incentive motivation" for the outcome (Dickinson and Balleine 2002; McClure et al. 2003; Berridge 2004) likely also involve temporal discounting (which we have not directly modeled here) by which delayed rewards are viewed as less valuable than proximal ones. The additional value of receiving a larger reward faster could thus be expected to offset the cost of a more vigorous response.

Future directions

In our model, we used response timing as a proxy for response vigor, adopting the rather simplistic view that animals select a particular latency for their chosen action and then set about doing it exactly that slowly and with no interruption. Although this has allowed us to capture some effects of response timing in free-operant tasks, the model in its current form misses others such as the prominent scalloped responding in fixed interval schedules (Gallistel and Gibbon 2000). A key lacuna in this respect is the assumption that animals can time latencies exactly, whereas it is known that interval timing is notoriously inaccurate (Gibbon 1977; Gallistel and Gibbon 2000). However, it is straightforward to include temporal noise in the model, and we expect our main conclusions about vigor to remain valid.

Among the most critical issues left for future work are aversively motivated conditioning and the relation between dopamine and serotonin in the striatum. Daw et al. (2002) suggested an opponency between serotonin and dopamine in controlling appetitive and aversive conditioning based on data showing various forms of antagonism between these neuromodulators (e.g., Fletcher and Korth 1999) and in light of long-standing psychological ideas (Konorski 1967; Solomon and Corbit 1974; Dickinson and Balleine 2002) that two opponent motivational systems exist. Their model suggested that if phasic dopamine reports appetitive prediction errors, then phasic serotonin should report aversive prediction errors. This was construed in the context of an average-reward RL model, rather like the one we have discussed here. Furthermore, Daw et al. (2002) suggested that opponency also extended to the tonic signals, with tonic dopamine representing the average rate of punishment (inspired by microdialysis data, suggesting dopamine concentrations rise during prolonged aversive stimulation) and tonic serotonin, conversely, reporting the average reward rate.

The present model's association of tonic dopamine with average reward rather than punishment seems to reverse this prior suggestion. However, it may be possible to integrate the two views. For instance, in active avoidance tasks, responding is known to be under dopaminergic control. In this case, there is an analogous form of opportunity cost that forces fast avoidance coming from the possibility of failing to escape a punishment. This link between average rate of punishment and vigor could potentially be realized by the same dopaminergic substrate as the appetitive energizing we have discussed. In any event, the hypothetical joint dopaminergic and serotonergic coding for appetitive and aversive signals presents burning empirical questions.

Conclusions

In conclusion, we suggest a computational model of striatal dopamine which incorporates both tonic and phasic dopamine signals into an action selection framework emphasizing both the identity of the chosen action and the vigor of its execution. The novel aspect of our model—the account of action vigor and its relation to the average reward rate, which we suggest is reported by tonic dopamine levels—allows for the modeling of free-operant behavior within the framework of RL and the understanding of effects of both motivational manipulations and dopaminergic interventions on response rates. Our account emphasizes the nonbinary nature of cost–benefit tradeoffs which animals and humans continuously face, as the decision on action vigor (or latency) embodies a continuous valued decision as to how much effort to exert given the available benefits. This framework provides another step in the direction of clarifying the role of striatal dopamine and its effects on behavior, even as it opens the door to a wealth of experimental work that can quantify the precise interplay between cost and benefit and between tonic and phasic dopamine.

Acknowledgements This work was funded by the Gatsby Charitable Foundation, a Hebrew University Rector Fellowship (Y.N.), the Royal Society (N.D.), and the EU Bayesian Inspired Brain and Artefacts (BIBA) project (N.D. and P.D.). We are grateful to Saleem Nicola, Mark Walton, and Matthew Rushworth for valuable discussions.

Appendix

Here we describe in more mathematical detail the proposed RL model of free-operant response rates (Niv et al. 2005a) from which the results in this paper were derived.

Formally, the action selection problem faced by the rat can be characterized by a series of states, $S \in \boldsymbol{S}$, in each of

which the rat must choose an action and a latency (a,τ) which will entail a unit cost, C_u, and a vigor cost, C_v/τ, and result in a possible transition to a new state, S', and a possible immediate reward with utility, U_r. The unit cost constant C_u and the vigor cost constant C_v can take different values depending on the identity of the currently chosen action $a\in${LP, NP, "Other"} and on that of the previously performed action. The transitions between states and the probability of reward for each action are governed by the schedule of reinforcement. For instance, in a random-ratio 5 (RR5) schedule, every LP action has p=0.2 probability of inducing a transition from the state in which no food is available in the magazine to that in which food is available. An NP action in the "no-reward-available" state is never rewarded and, conversely, is rewarded with certainty (p_r=1) in the "food-available-in-magazine" state. As a simplification, for each reinforcement schedule, we define states that incorporate all the available information relevant to decision making, such as the identity of the previously chosen action, whether or not food is available in the magazine, the time that has elapsed since the last lever press (in random-interval schedules only), and the number of lever presses since the last reward (in fixed ratio schedules only). The animal's behavior in the experiment is thus fully described by the successive actions and latencies chosen at the different states the animal encountered {(a_i, τ_i, S_i), i=1,2,3, ...}. The average reward rate $\overline{R}$ is simply the sum of all the rewards obtained minus all the costs incurred, all divided by the total amount of time.

Using this formulation, we can define the differential value of a state, denoted $V(S)$, as the expected sum of future rewards minus costs encountered from this state and onward compared with the expected average reward rate. Defining the value as an expectation over a sum means that the value can be written recursively as the expected reward minus cost due to the current action, compared with the immediately forfeited average reward, plus the value of the next state (averaged over the possible next states). To find the optimal differential values of the different states, that is, the values $V^*(S)$ (and average value $\overline{R}^*$) given the optimal action selection strategy, we can simultaneously solve the set of equations defining these values:

$$V^*(S) = \max_{a,\tau}\left\{p_r(a,\tau,S)U_r - C_u\left(a,a_{prev}\right) - \frac{C_v\left(a,a_{prev}\right)}{\tau} - \overline{R}^*\cdot\tau + \sum_{S'\in\boldsymbol{S}} p(S'|a,\tau,S)V^*(S')\right\}, \quad (1)$$

in which there is one equation for every state $S\in\boldsymbol{S}$, and $p(S'|a,\tau,S)$ is the schedule-defined probability to transition to state S' given (a,τ) was performed at state S.

The theory of dynamic programming (Bertsekas and Tsitsiklis 1996) ensures that these equations have one solution for the optimal attainable average reward $\overline{R}^*$, and the optimal differential state values $V^*(S)$ (which are defined up to an additive constant). This solution can be found using iterative dynamic programming methods such as "value iteration" (Bertsekas and Tsitsiklis 1996) or approximated through online sampling of the task dynamics and temporal-difference learning (Schwartz 1993; Mahadevan 1996; Sutton and Barto 1998). Here we used the former and report results using the true optimal differential values. We compare these model results to the steady-state behavior of well-trained animals as the optimal values correspond to values learned online throughout an extensive training period.

Given the optimal state values, the optimal differential value of an (a,τ) pair taken at state S, denoted $Q^*(a,\tau,S)$, is simply:

$$Q^*(a,\tau,S) = p_r(a,\tau,S)\cdot U_r - C_u\left(a,a_{prev}\right) - \frac{C_v\left(a,a_{prev}\right)}{\tau} - \overline{R}^*\cdot\tau + \sum_{S'\in\boldsymbol{S}} p(S'|a,\tau,S)V^*(S') \quad (2)$$

The animal can select actions optimally (that is, such as to obtain the maximal possible average reward rate $\overline{R}^*$) by comparing the differential values of the different (a,τ) pairs at the current state and choosing the action and latency that have the highest value. Alternatively, to allow more flexible behavior and occasional exploratory actions (Daw et al. 2006), response selection can be based on the so-called "soft-max" rule (or Boltzmann distribution) in which the probability of choosing an (a,τ) pair is proportional to its differential value. In this case, which is the one we used here, actions that are "almost optimal" are chosen almost as frequently as actions that are strictly optimal. Specifically, the probability of choosing (a,τ) in state S is:

$$p(a,\tau,S) = \frac{e^{\beta Q^*(a,\tau,S)}}{\sum_{a',\tau'} e^{\beta Q^*(a',\tau',S)}}, \quad (3)$$

where β is the inverse temperature controlling the steepness of the soft-max function (a value of zero corresponds to uniform selection of actions, whereas higher values correspond to a more maximizing strategy).

To simulate the (immediate) effects of depletion of tonic dopamine (Fig. 3b), Q values were recomputed from the optimal V values (using Eq. 2), but taking into account a lower average reward rate (specifically, $\overline{R}_{depleted} = 0.4\overline{R}^*$). Actions were then chosen as usual, using the soft-max function of these new Q values, to generate behavior.

Finally, note that Eq. 2 is a function relating actions and latencies to values. Accordingly, one way to find the optimal latency is to differentiate Eq. 2 with respect to τ and find its maximum. For ratio schedules (in which the identity and value of the subsequent state S' is not dependent on τ), this gives:

$$\tau^* = \sqrt{\frac{C_v}{\overline{R}^*}}, \tag{4}$$

showing that the optimal latency τ^* depends solely on the vigor cost constant and the average reward rate. This is true regardless of the action a chosen, which is why a change in the average reward has a similar effect on the latencies of all actions. In interval schedules, the situation is slightly more complex because the identity of the subsequent state is dependent on the latency, and this must be taken into account when taking the derivative. However, in this case as well, the optimal latency is inversely related to the average reward rate.

References

Aberman JE, Salamone JD (1999) Nucleus accumbens dopamine depletions make rats more sensitive to high ratio requirements but do not impair primary food reinforcement. Neuroscience 92 (2):545–552

Ainslie G (1975) Specious reward: a behavioural theory of impulsiveness and impulse control. Psychol Bull 82:463–496

Barrett JE, Stanley JA (1980) Effects of ethanol on multiple fixed-interval fixed-ratio schedule performances: dynamic interactions at different fixed-ratio values. J Exp Anal Behav 34 (2):185–198

Barto AG (1995) Adaptive critics and the basal ganglia. In: Houk JC, Davis JL, Beiser DG (eds) Models of information processing in the basal ganglia. MIT Press, Cambridge, pp 215–232

Beninger RJ (1983) The role of dopamine in locomotor activity and learning. Brain Res Brain Res Rev 6:173–196

Bergstrom BP, Garris PA (2003) 'Passive stabilization' of striatal extracellular dopamine across the lesion spectrum encompassing the presymptomatic phase of Parkinson's disease: a voltammetric study in the 6-OHDA lesioned rat. J Neurochem 87(5):1224–1236

Berridge KC (2004) Motivation concepts in behavioral neuroscience. Physiol Behav 81(2):179–209

Berridge KC, Robinson TE (1998) What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? Brain Res Brain Res Rev 28:309–369

Bertsekas DP, Tsitsiklis JN (1996) Neuro-dynamic programming. Athena, Belmont

Bolles RC (1967) Theory of motivation. Harper and Row, New York

Carr GD, White NM (1987) Effects of systemic and intracranial amphetamine injections on behavior in the open field: a detailed analysis. Pharmacol Biochem Behav 27:113–122

Catania AC, Reynolds GS (1968) A quantitative analysis of the responding maintained by interval schedules of reinforcement. J Exp Anal Behav 11:327–383

Catania AC, Matthews TJ, Silverman PJ, Yohalem R (1977) Yoked variable-ratio and variable-interval responding in pigeons. J Exp Anal Behav 28:155–161

Chéramy A, Barbeito L, Godeheu G, Desce J, Pittaluga A, Galli T, Artaud F, Glowinski J (1990) Respective contributions of neuronal activity and presynaptic mechanisms in the control of the in vivo release of dopamine. J Neural Transm Suppl 29:183–193

Chesselet MF (1990) Presynaptic regulation of dopamine release. Implications for the functional organization of the basal ganglia. Ann N Y Acad Sci 604:17–22

Correa M, Carlson BB, Wisniecki A, Salamone JD (2002) Nucleus accumbens dopamine and work requirements on interval schedules. Behav Brain Res 137:179–187

Cousins MS, Atherton A, Turner L, Salamone JD (1996) Nucleus accumbens dopamine depletions alter relative response allocation in a T-maze cost/benefit task. Behav Brain Res 74:189–197

Daw ND (2003) Reinforcement learning models of the dopamine system and their behavioral implications. Unpublished doctoral dissertation, Carnegie Mellon University

Daw ND, Touretzky DS (2002) Long-term reward prediction in TD models of the dopamine system. Neural Comp 14:2567–2583

Daw ND, Kakade S, Dayan P (2002) Opponent interactions between serotonin and dopamine. Neural Netw 15(4–6):603–616

Daw ND, Niv Y, Dayan P (2005) Uncertainty based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8(12):1704–1711

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879

Dawson GR, Dickinson A (1990) Performance on ratio and interval schedules with matched reinforcement rates. Q J Exp Psychol B 42:225–239

Denk F, Walton ME, Jennings KA, Sharp T, Rushworth MF, Bannerman DM (2005) Differential involvement of serotonin and dopamine systems in cost–benefit decisions about delay or effort. Psychopharmacology (Berl) 179(3):587–596

Dickinson A (1985) Actions and habits: the development of behavioural autonomy. Philos Trans R Soc Lond B Biol Sci 308(1135):67–78

Dickinson A, Balleine B (1994) Motivational control of goal-directed action. Anim Learn Behav 22:1–18

Dickinson A, Balleine B (2002) The role of learning in the operation of motivational systems. In: Pashler H, Gallistel R (eds) Stevens' handbook of experimental psychology. Learning, motivation and emotion, 3rd edn, vol 3. Wiley, New York, pp 497–533

Dickinson A, Smith J, Mirenowicz J (2000) Dissociation of Pavlovian and instrumental incentive learning under dopamine agonists. Behav Neurosci 114(3):468–483

Domjan M (2003) Principles of learning and behavior, 5th edn. Thomson/Wadsworth, Belmont

Dragoi V, Staddon JER (1999) The dynamics of operant conditioning. Psychol Rev 106(1):20–61

Evenden JL, Robbins TW (1983) Increased dopamine switching, perseveration and perseverative switching following D-amphetamine in the rat. Psychopharmacology (Berl) 80:67–73

Faure A, Haberland U, Condé F, Massioui NE (2005) Lesion to the nigrostriatal dopamine system disrupts stimulus–response habit formation. J Neurosci 25:2771–2780

Fiorillo C, Tobler P, Schultz W (2003) Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299 (5614):1898–1902

Fletcher PJ, Korth KM (1999) Activation of 5-HT1B receptors in the nucleus accumbens reduces amphetamine-induced enhancement of responding for conditioned reward. Psychopharmacology (Berl) 142:165–174

Floresco SB, West AR, Ash B, Moore H, Grace AA (2003) Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. Nat Neurosci 6(9):968–973

Foster TM, Blackman KA, Temple W (1997) Open versus closed economies: performance of domestic hens under fixed-ratio schedules. J Exp Anal Behav 67:67–89

Friston KJ, Tononi G, Reeke GNJ, Sporns O, Edelman GM (1994) Value-dependent selection in the brain: simulation in a synthetic neural model. Neuroscience 59(2):229–243

Gallistel CR, Gibbon J (2000) Time, rate and conditioning. Psychol Rev 107:289–344

Gallistel CR, Stellar J, Bubis E (1974) Parametric analysis of brain stimulation reward in the rat: I. The transient process and the memory-containing process. J Comp Physiol Psychol 87:848–860

Gibbon J (1977) Scalar expectancy theory and Weber's law in animal timing. Psychol Rev 84(3):279–325

Goto Y, Grace A (2005) Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. Nat Neurosci 8:805–812

Grace AA (1991) Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. Neuroscience 41(1):1–24

Hernandez G, Hamdani S, Rajabi H, Conover K, Stewart J, Arvanitogiannis A, Shizgal P (2006) Prolonged rewarding stimulation of the rat medial forebrain bundle: neurochemical and behavioral consequences. Behav Neurosci 120(4):888–904

Herrnstein RJ (1961) Relative and absolute strength of response as a function of frequency of reinforcement. J Exp Anal Behav 4 (3):267–272

Herrnstein RJ (1970) On the law of effect. J Exp Anal Behav 13 (2):243–266

Houk JC, Adams JL, Barto AG (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Houk JC, Davis JL, Beiser DG (eds) Models of information processing in the basal ganglia. MIT Press, Cambridge, pp 249–270

Ikemoto S, Panksepp J (1999) The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. Brain Res Brain Res Rev 31:6–41

Jackson DM, Anden N, Dahlstrom A (1975) A functional effect of dopamine in the nucleus accumbens and in some other dopamine-rich parts of the rat brain. Psychopharmacologia 45:139–149

Kacelnik A (1997) Normative and descriptive models of decision making: time discounting and risk sensitivity. In: Bock GR, Cardew G (eds) Characterizing human psychological adaptations: Ciba Foundation symposium 208. Wiley, Chichester, pp 51–70

Killeen PR (1995) Economics, ecologies and mechanics: the dynamics of responding under conditions of varying motivation. J Exp Anal Behav 64:405–431

Konorski J (1967) Integrative activity of the brain: an interdisciplinary approach. University of Chicago Press, Chicago

Lauwereyns J, Watanabe K, Coe B, Hikosaka O (2002) A neural correlate of response bias in monkey caudate nucleus. Nature 418 (6896):413–417

Le Moal M, Simon H (1991) Mesocorticolimbic dopaminergic network: functional and regulatory roles. Physiol Rev 71:155–234

Ljungberg T, Enquist M (1987) Disruptive effects of low doses of D-amphetamine on the ability of rats to organize behaviour into functional sequences. Psychopharmacology (Berl) 93:146–151

Ljungberg T, Apicella P, Schultz W (1992) Responses of monkey dopaminergic neurons during learning of behavioral reactions. J Neurophys 67:145–163

Lodge DJ, Grace AA (2005) The hippocampus modulates dopamine neuron responsivity by regulating the intensity of phasic neuron activation. Neuropsychopharmacology 31:1356–1361

Lodge DJ, Grace AA (2006) The laterodorsal tegmentum is essential for burst firing of ventral tegmental area dopamine neurons. Proc Nat Acad Sci U S A 103(13):5167–5172

Lyon M, Robbins TW (1975) The action of central nervous system stimulant drugs: a general theory concerning amphetamine effects. In: Current developments in psychopharmacology. Spectrum, New York, pp 80–163

Mahadevan S (1996) Average reward reinforcement learning: foundations, algorithms and empirical results. Mach Learn 22:1–38

Mazur JA (1983) Steady-state performance on fixed-, mixed-, and random-ratio schedules. J Exp Anal Behav 39(2):293–307

McClure SM, Daw ND, Montague PR (2003) A computational substrate for incentive salience. Trends Neurosci 26(8):423–428

Mingote S, Weber SM, Ishiwari K, Correa M, Salamone JD (2005) Ratio and time requirements on operant schedules: effort-related effects of nucleus accumbens dopamine depletions. Eur J Neurosci 21:1749–1757

Montague PR (2006) Why choose this book?: how we make decisions. Dutton, New York

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16(5):1936–1947

Moore H, West AR, Grace AA (1999) The regulation of forebrain dopamine transmission: relevance to the psychopathology of schizophrenia. Biol Psychiatry 46:40–55

Murschall A, Hauber W (2006) Inactivation of the ventral tegmental area abolished the general excitatory influence of Pavlovian cues on instrumental performance. Learn Mem 13:123–126

Niv Y, Daw ND, Dayan P (2005a) How fast to work: response vigor, motivation and tonic dopamine. In: Weiss Y, Schölkopf B, Platt J (eds) NIPS 18. MIT Press, Cambridge, pp 1019–1026

Niv Y, Daw ND, Joel D, Dayan P (2005b) Motivational effects on behavior: towards a reinforcement learning model of rates of responding. COSYNE 2005, Salt Lake City

Niv Y, Joel D, Dayan P (2006) A normative perspective on motivation. Trends Cogn Sci 10:375–381

Oades RD (1985) The role of noradrenaline in tuning and dopamine in switching between signals in the CNS. Neurosci Biobehav Rev 9 (2):261–282

Packard MG, Knowlton BJ (2002) Learning and memory functions of the basal ganglia. Annu Rev Neurosci 25:563–593

Phillips PEM, Wightman RM (2004) Extrasynaptic dopamine and phasic neuronal activity. Nat Neurosci 7:199

Phillips PEM, Stuber GD, Heien MLAV, Wightman RM, Carelli RM (2003) Subsecond dopamine release promotes cocaine seeking. Nature 422:614–618

Redgrave P, Prescott TJ, Gurney K (1999) The basal ganglia: a vertebrate solution to the selection problem? Neuroscience 89:1009–1023

Robbins TW, Everitt BJ (1982) Functional studies of the central catecholamines. Int Rev Neurobiol 23:303–365

Roitman MF, Stuber GD, Phillips PEM, Wightman RM, Carelli RM (2004) Dopamine operates as a subsecond modulator of food seeking. J Neurosci 24(6):1265–1271

Salamone JD, Correa M (2002) Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. Behav Brain Res 137:3–25

Salamone JD, Wisniecki A, Carlson BB, Correa M (2001) Nucleus accumbens dopamine depletions make animals highly sensitive to high fixed ratio requirements but do not impair primary food reinforcement. Neuroscience 5(4):863–870

Satoh T, Nakai S, Sato T, Kimura M (2003) Correlated coding of motivation and outcome of decision by dopamine neurons. J Neurosci 23(30):9913–9923

Schoenbaum G, Setlow B, Nugent S, Saddoris M, Gallagher M (2003) Lesions of orbitofrontal cortex and basolateral amygdala complex

disrupt acquisition of odor-guided discriminations and reversals. Learn Mem 10:129–140
Schultz W (1998) Predictive reward signal of dopamine neurons. J Neurophys 80:1–27
Schultz W, Apicella P, Ljungberg T (1993) Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J Neurosci 13:900–913
Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599
Schwartz A (1993) A reinforcement learning method for maximizing undiscounted rewards. In: Proceedings of the tenth international conference on machine learning. Morgan Kaufmann, San Francisco, pp 298–305
Sokolowski JD, Salamone JD (1998) The role of accumbens dopamine in lever pressing and response allocation: effects of 6-OHDA injected into core and dorsomedial shell. Pharmacol Biochem Behav 59(3):557–566
Solomon RL, Corbit JD (1974) An opponent-process theory of motivation. I. Temporal dynamics of affect. Psychol Rev 81:119–145
Staddon JER (2001) Adaptive dynamics. MIT Press, Cambridge
Sutton RS, Barto AG (1981) Toward a modern theory of adaptive networks: expectation and prediction. Psychol Rev 88:135–170
Sutton RS, Barto AG (1990) Time-derivative models of Pavlovian reinforcement. In: Gabriel M, Moore J (eds) Learning and computational neuroscience: foundations of adaptive networks. MIT Press, Cambridge, pp 497–537
Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge
Taghzouti K, Simon H, Louilot A, Herman J, Le Moal M (1985) Behavioral study after local injection of 6-hydroxydopamine into the nucleus accumbens in the rat. Brain Res 344:9–20
Takikawa Y, Kawagoe R, Itoh H, Nakahara H, Hikosaka O (2002) Modulation of saccadic eye movements by predicted reward outcome. Exp Brain Res 142(2):284–291
Taylor JR, Robbins TW (1984) Enhanced behavioural control by conditioned reinforcers following microinjections of D-amphetamine into the nucleus accumbens. Psychopharmacology (Berl) 84:405–412
Taylor JR, Robbins TW (1986) 6-Hydroxydopamine lesions of the nucleus accumbens, but not of the caudate nucleus, attenuate enhanced responding with reward-related stimuli produced by intra-accumbens D-amphetamine. Psychopharmacology (Berl) 90:390–397
Tobler P, Fiorillo C, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. Science 307(5715):1642–1645
van den Bos R, Charria Ortiz GA, Bergmans AC, Cools AR (1991) Evidence that dopamine in the nucleus accumbens is involved in the ability of rats to switch to cue-directed behaviours. Behav Brain Res 42:107–114
Waelti P, Dickinson A, Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. Nature 412:43–48
Walton ME, Kennerley SW, Bannerman DM, Phillips PEM, Rushworth MFS (2006) Weighing up the benefits of work: behavioral and neural analyses of effort-related decision making. Neural networks (in press)
Watanabe M, Cromwell H, Tremblay L, Hollerman J, Hikosaka K, Schultz W (2001) Behavioral reactions reflecting differential reward expectations in monkeys. Exp Brain Res 140(4):511–518
Weiner I (1990) Neural substrates of latent inhibition: the switching model. Psychol Bull 108:442–461
Weiner I, Joel D (2002) Dopamine in schizophrenia: dysfunctional information processing in basal ganglia-thalamocortical split circuits. In: Chiara GD (ed) Handbook of experimental pharmacology, vol 154/II. Dopamine in the CNS II. Springer, Berlin Heidelberg New York, pp 417–472
Wickens J (1990) Striatal dopamine in motor activation and reward-mediated learning: steps towards a unifying model. J Neural Transm 80:9–31
Wickens J, Kötter R (1995) Cellular models of reinforcement. In: Houk JC, Davis JL, Beiser DG (eds) Models of information processing in the basal ganglia. MIT Press, Cambridge, pp 187–214
Wilson C, Nomikos GG, Collu M, Fibiger HC (1995) Dopaminergic correlates of motivated behavior: importance of drive. J Neurosci 15(7):5169–5178
Wise RA (2004) Dopamine, learning and motivation. Nat Rev Neurosci 5:483–495
Yin HH, Knowlton BJ, Balleine BW (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci 19:181–189
Zuriff GE (1970) A comparison of variable-ratio and variable-interval schedules of reinforcement. J Exp Anal Behav 13:369–374