Computational roles for dopamine in behavioural control

P. Read Montague[1,2], Steven E. Hyman[3] & Jonathan D. Cohen[4,5]

[1]Department of Neuroscience and [2]Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA (e-mail: read@bcm.tmc.edu)
[3]Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: seh@harvard.edu)
[4]Department of Psychiatry, University of Pittsburgh and [5]Department of Psychology, Center for the Study of Brain, Mind & Behavior, Green Hall, Princeton University, Princeton, New Jersey 08544, USA (e-mail: jdc@princeton.edu)

Neuromodulators such as dopamine have a central role in cognitive disorders. In the past decade, biological findings on dopamine function have been infused with concepts taken from computational theories of reinforcement learning. These more abstract approaches have now been applied to describe the biological algorithms at play in our brains when we form value judgements and make choices. The application of such quantitative models has opened up new fields, ripe for attack by young synthesizers and theoreticians.

The concept of behavioural control is intimately tied to the valuation of resources and choices. For example, a creature that moves left instead of right may forgoe the food and other resources that it could have obtained had it chosen right. Such stark, yet simple economic realities select for creatures that evaluate the world quickly and choose appropriate behaviour based on those valuations. From the point of view of selection, the most effective valuations are those that improve reproductive success. This prescription for valuation yields a formula for desires or goals: an organism should desire those things deemed most valuable to it. All mobile organisms possess such discriminatory capacities and can rank numerous dimensions in their world along axes that extend from good to bad. A kind of facile biological wisdom is built into these simple observations and we should expect valuation mechanisms to be built into our nervous systems at every level, from the single neuron to the decision algorithms used in complex social settings.

These ideas have recently been upgraded from provocative biological musings to real computational models of how the nervous system sets goals, computes values of particular resources or options, and uses both to guide sequences of behavioural choices. Such models have cast as important players our midbrain's dopamine neurons, whose actions define 'rewards' — our goals or desires — that should be sought. These neurons have a central role in guiding our behaviour and thoughts. They are hijacked by every addictive drug; they malfunction in mental illness; and they are lost in dramatically impairing illnesses such as Parkinson's disease. If dopamine systems are overstimulated, we may hear voices, experience elaborate bizarre cognitive distortions, or engage excessively in dangerous goal-directed behaviour. Dopamine function is also central to the way that we value our world, including the way that we value money and other human beings.

The full story of behavioural control requires vastly more than simple models of dopaminergic function. But here we show how one branch of computational theory — reinforcement learning — has informed both the design and interpretation of experiments that probe how the dopamine system influences sequences of choices made about rewards. These models are maturing rapidly and may even guide our understanding of other neuromodulatory systems in the brain, although such applications are still in their infancy.

Reinforcement signals define an agent's goals

Reinforcement learning theories seek to explain how organisms learn to organize their behaviour under the influence of rewards[1]. 'Reward' is an old psychological term defined by Merriam Webster's dictionary as "a stimulus administered to an organism following a correct or desired response that increases the probability of occurrence of the response". Here, we show that current theories of reinforcement learning provide a formal framework for connecting the physiological actions of specific neuromodulatory systems to behavioural control. We focus on dopaminergic systems primarily because they have been most extensively modelled and because they play a major role in decision-making, motor output, executive control and reward-dependent learning[2–5]. We show how the dopaminergic models provide a way to understand neuroimaging experiments on reward expectancy and cognitive control in human subjects. Finally, we suggest that this same class of model has matured sufficiently for it to be used to address important disturbances in neuromodulation associated with many psychiatric disorders.

Despite its name, reinforcement learning is not simply a modern recapitulation of stimulus–response learning, familiar from the classical and instrumental conditioning literature[6]. Traditional stimulus–response models focused on how direct associations can be learned between stimuli and responses, overlooking the possibility that numerous internal states intervene between the stimulus and its associated response. However, animals clearly have covert internal states that affect overt, measurable behaviour. Reinforcement learning theory explicitly models such intervening states, assumes that some are more desirable than others, and asks how do animals learn to achieve desired states and avoid undesirable ones as efficiently as possible? The answer to this question shows how reinforcement signals define an agent's goals. For simplicity, we focus only on rewards. However, the same story can be told using negative reinforcers (punishments).

We refer to the state engendered by a reward as a 'goal'. Goals can exist at numerous levels and direct behaviour over many timescales. Goals for humans range from the most basic (for example, procuring something to eat in the next minute) to the most abstract and complex (such as planning a career). In reinforcement learning, it is assumed that the fundamental goal of the agent (learner) is to learn to take actions that are most likely to lead to the greatest accrual of rewards in the future. This goal is achieved under the guidance of simple scalar quantities called reinforcement signals. These signals

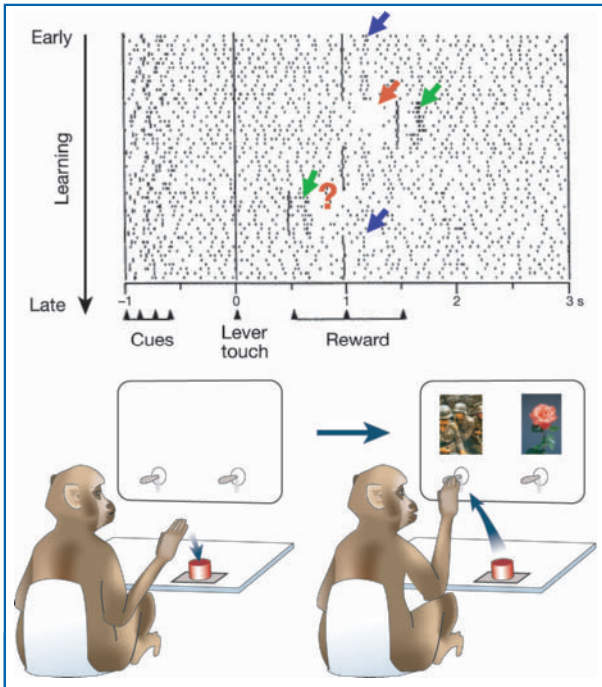

Figure 1 TD prediction-error signal encoded in dopamine neuron firing. Electrophysiological recordings from a single dopamine neuron in a monkey during reward-dependent discrimination task. The animal presses a key, two pictures are presented, the animal releases the key and hits the lever under the rewarded picture. If a correct choice is made, juice is delivered after a fixed delay. Juice delivery is marked by vertical bars; neuron spikes by dots. Early on, the juice delivery causes a burst response (top blue arrowhead). This 'surprise' response diminishes to zero by the end of learning (bottom blue arrowhead). A 'catch trial' using a (surprising) delayed reward time exposes the typical pause (red arrowhead) and burst (top green arrowhead) response. The pause signals that 'things are worse than expected' and the burst signals that 'things are better than expected'. In the second catch trial, the reward is again surprising, but early rather than late. The burst response for the new delivery time is apparent (lower green arrowhead), but the pause response is less certain (red question mark). Adapted from ref. 13.

serve to criticize specific actions or contemplated actions with respect to how effectively they serve the agent's goals. In reinforcement learning, one common goal is the maximization of total future reward[6].

Every reinforcement learning system possesses three explicitly implemented components: (1) a 'reinforcement signal' that assigns a numerical quantity to every state of the agent. Reinforcement signals can be negative or positive. They define the agent's immediate goals by reporting on what is good or bad 'right now'; (2) a stored 'value function' that formalizes the idea of longer-term judgments by assigning a 'value' to the current state of the agent (see Box 1); (3) a 'policy function' that maps the agent's states to its actions. Policies are typically stochastic: they assign a probability to each possible action that can be taken from the current state, with the probability weighted by the value of the next state produced by that action.

A more concrete description reads as iterations of the following recipe: (1) organism is in state X and receives reward information; (2) organism queries stored value of state X; (3) organism updates stored value of state X based on current reward information; (4) organism selects action based on stored policy; and (5) organism transitions to state Y and receives reward information.

In one form of reinforcement learning called temporal-difference learning, a critical signal is the reward-prediction error (also called the temporal-difference, or TD error)[7–9]. Unlike the well-known psychological learning rule proposed by Rescorla and Wagner[10] in 1972, this error function is not simply a difference between the received reward and predicted reward; instead, it incorporates information about the next prediction made by the reward-prediction system[11]. In words: current TD error = current reward + γ·next prediction − current prediction. Here, the words 'current' and 'next' refer respectively to the present state and to the subsequent state of the learner; γ is a factor between 0 and 1 that weights the relative influence of the next prediction. By using this reward-prediction error to refine predictions of reward for each state, the system can improve its estimation of the value of each state, and improve its policy function's ability to choose actions that lead to more reward.

The reward-prediction-error hypothesis

Over the past decade, experimental work by Wolfram Schultz and colleagues has shown that dopaminergic neurons of the ventral tegmental area and substantia nigra show phasic changes in spike activity that correlate with the history of reward delivery[12–16]. It was proposed that these phasic activity changes encode a 'prediction error about summed future reward' (as described above): this hypothesis has been tested successfully against a range of physiological data[2–3]. The 'pause' and 'burst' responses of dopamine neurons that support a reward-prediction-error hypothesis are shown in Fig. 1. The bursts signal a positive reward-prediction error ('things are better than expected'), and the pauses signal a negative prediction error ('things are worse than expected'). Activity that remains close to the baseline signals that 'things are just as expected'. However, this verbal interpretation of dopaminergic activity belies the sophistication of the underlying neural computations[1] (Box 1).

Value binding and incentive salience

We have presented theoretical evidence that phasic bursts and pauses in midbrain dopaminergic activity are consistent with the formal construct of a reward-prediction error used by reinforcement learning systems (Fig. 1; Box 1). This interpretation is consistent with a long history of physiological and pharmacological data showing that dopamine is involved in appetitive approach behaviour[17–19], and is a key component in the pathologies of behavioural control associated with drug addiction[20–21].

One finding offered as a challenge to the models discussed so far is that antagonism of dopamine receptors does not change the appetitive value of food rewards but does prevent the treated animal from initiating actions that allow it to obtain the food reward[17,22]. In these experiments, animals treated with dopamine-receptor blockers are virtually unable to link sequences of actions to obtain a food reward, but they will consume the same amount as untreated animals if they are moved close to the food rewards by the experimenter (Fig. 2). This conclusion also holds for the inhibition of dopamine neuron firing by gamma-aminobutyric acid (GABA) injected directly into the ventral tegmental area (Fig. 2). These data suggest that interfering with dopamine transmission does not alter the internal evaluation of rewards, but simply the ability to act on those valuations. Addressing these data at a conceptual level, Berridge and Robinson have proposed that dopamine mediates the 'binding' between the hedonic evaluation of stimuli and the assignment of these values to objects or acts[17]. They call this idea 'incentive salience'. Although competing psychological explanations differ with respect to the specific claims of incentive salience[19,23,24], they all agree that dopamine release and binding is a necessary link between the evaluation of potential future rewards and the policy (sequence of actions) that acquires the rewards. Here, we refer to this link as value binding and distinguish three components: (1) the value computation; (2) the link to a policy (value binding); and (3) execution of the policy.

Incentive salience and actor–critic models

There is a class of reinforcement learning model, called the actor–critic that is closely related to the Berridge and Robinson model for the role of dopamine in value and action learning[1,9]. In these models, the 'critic' carries the reward-prediction error associated

Box 1

Value functions and prediction errors

The value function

In the simplest TD models of dopamine systems, the reward-prediction error depends on a value function that equates the value V of the current state s at time t with the average sum of future rewards received up until the end of a learning trial.

$$\begin{aligned} V(s_t) &= \text{average sum of future rewards delivered from state } s_t \text{ until the end of a learning trial} \\ &= \text{average } [r_t + r_{t+1} + r_{t+2} + \ldots + r \text{ (trial's end)}] \\ &= E\left[\sum_{\tau \in Trial} r(\tau)\right] \end{aligned} \quad (1)$$

E is the expected value operator. There are two sources of randomness over which the above averaging occurs. First, the rewards in a trial $[r_t + r_{t+1} + r_{t+2} + \ldots + r \text{ (trial's end)}]$ are random variables indexed by the time t. For example, r_{t+2} is a sample of the distribution of rewards received two timesteps into the trial. The idea is that the animal can learn the average value of these rewards by repeating learning trials, and by revisiting state s_t sufficiently frequently for its nervous system to be able to estimate the average value of each of the rewards received from state s_t until the end of the trial. The second source of randomness is the probabilistic transition from one state at time t to a succeeding state s_{t+1} at a later time $t+1$. The value function, stored within the nervous system of the creature, provides an assessment of the likely future rewards for each state of the creature; that is, the value must somehow be associated with the state. However, as written in equation (1), it would be virtually impossible to make good estimates of the ideal $V(s_t)$ as it is now defined. This is because the creature would have to wait until all rewards were received within a trial before deciding on the value of its state at the beginning of the trial. By that time, it is too late for such a computation to be useful. This problem becomes worse in real-world settings. Fortunately, equation (1) provides a way out of this dilemma because it obeys a recursion relation through time:

$$V(s_t) = E[r_t] + V(s_{t+1}) \quad (2)$$

This recursion relation shows that information about the value of a state s_t is available using only the value $V(s_t)$ of the current state s_t and the value of its successor state s_{t+1}. Until this point, we have been discussing the ideal case for V. However, as indicated above, V cannot be known exactly in the real world. Instead, an estimate $\hat{V}$ of V must be formed within the nervous system. The TD algorithm learns an approximation $\hat{V}$ of the value function V. It uses a prediction-error signal:

$$\begin{aligned} \delta(t) &= \text{prediction error } (t) = E[r_t] + \hat{V}(s_{t+1}) - \hat{V}(s_t) \\ &\approx \text{current reward} + \text{next prediction} - \text{current prediction} \end{aligned} \quad (3)$$

This TD error signal reproduces the phasic burst and pause responses measured in dopamine neurons recorded in alert monkeys during learning tasks. The next value of each adaptable weight $w(t+1)$ used to estimate V is incremented or decremented in proportion to the product of the current prediction error $\delta(t)$ and the current representation $s(t)$ of the stimulus responsible for the prediction.

$$w(t+1) = w(t) + \lambda s(t) \cdot \delta(t) \quad (4)$$

Here, λ is a learning rate.

Exponential discounting of future rewards

The artificial truncation at the end of a trial (equation (1)) can be handled theoretically in several ways. One popular formalization is to weight the near future more than the distant future. In this case, the analogue to equation (1) takes the form:

$$\begin{aligned} V_d(\delta(t)) &= \text{average sum of discounted future rewards} \\ &= \text{average } [\gamma^0 r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2) + \ldots] \text{ for } 0 < \gamma < 1 \\ &= E\left[\sum_{\tau \geq t} \gamma^{\tau - t} r(\tau)\right] \end{aligned}$$

Using this weighted version of the value function, the learning episodes for a creature do not have to be artificially divided into 'trials'. An analogous reward-prediction-error signal can be formed and used in the same manner as above:

$$\begin{aligned} \delta(t) &= \text{prediction error } (t) = E[r_t] + \gamma \cdot \hat{V}(s_{t+1}) - \hat{V}(s_t) \\ &\approx \text{current reward} + \gamma \cdot \text{next prediction} - \text{current prediction} \end{aligned} \quad (5)$$

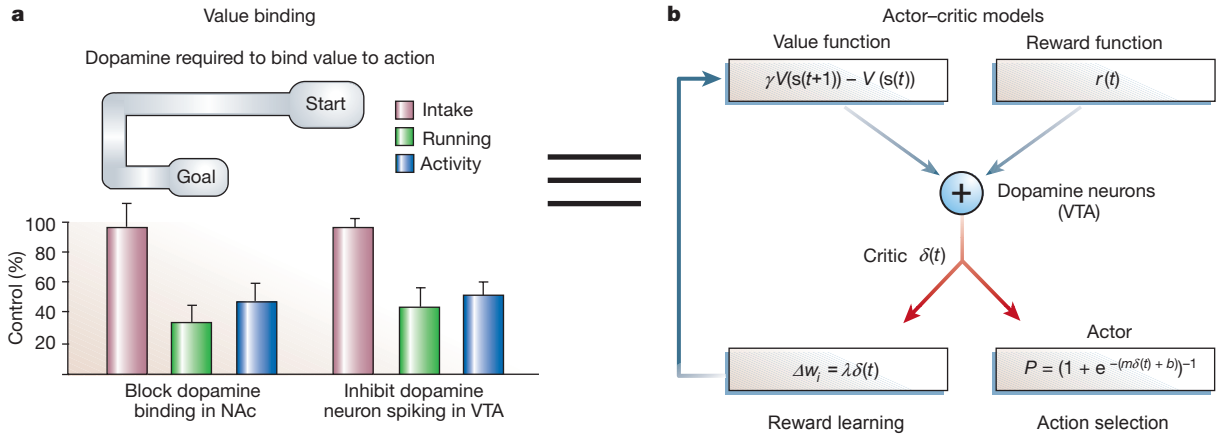

Figure 2 Equating incentive salience with the actor–critic model. **a**, Rats are trained to run a maze to acquire sugary water. If dopaminergic spiking is blocked (left histograms) in the VTA, then rats will generally not run down the maze to get a reward and are less active. However, if the experimenter moves them to the sugary water, the rats drink exactly the same amount as untreated rats. This suggests that the (hedonic) value of the sugary water has been computed but that the capacity to bind this value to actions required to obtain the water fails to function. The same effect results if dopamine's interaction with its receptor is blocked in an important downstream target of dopamine projections (right histograms). Adapted from refs 22 and 25. **b**, Actor–critic models use dopamine-encoded prediction-error signal in two roles: (1) to learn stimulus–reward associations, and (2) to assess actions or contemplated actions (notations are as in Box 1). Adapted from refs 2, 25, 83.

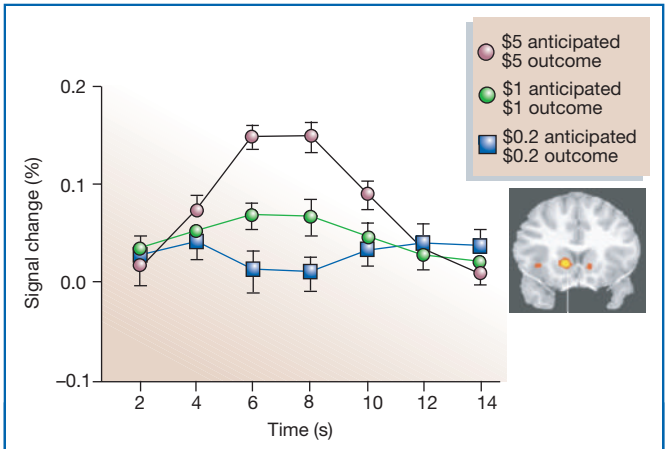

Figure 3 Scaled responses to a monetary reward in the ventral striatum. Action is required to receive a reward. The haemodynamic response is modulated by the amount of money received. In both cases, positive deviations in expectations make the responses bigger. Adapted from ref. 38.

with the states of the organism. The 'actor' uses this signal, or a closely related one, to learn stimulus–action associations, so that actions associated with higher rewards are more likely to be chosen. Together, these two components capture many features of the way that animals learn basic contingencies between their actions and the rewards associated with those actions. The original hypothesis concerning the role of dopamine in reinforcement learning proposed just such a dual use of the reward-prediction-error signal[2,25]. McClure and colleagues recently extended this original learning hypothesis to address the Berridge and Robinson model[26]. This work suggests a formal relationship between the incentive-salience ideas of Berridge and Robinson and actor–critic models in which incentive salience is equivalent to the idea of expected future value formalized in reinforcement learning models (Fig. 2).

Actor–critic models are now being used to address detailed issues concerning stimulus–action learning[8]. For example, extensions to actor–critic models have addressed the difference between learning goal-directed approach behaviour and learning automatic actions (habits), such as licking. There are several behavioural settings that support the contention that habit learning is handled by different neural systems from those involved in goal-directed learning[27,28]. Dayan and Balleine have recently offered a computational extension to actor–critic models to take account of this fact[29].

Rewards, critics and actors in the human brain

Recent functional magnetic resonance imaging (fMRI) experiments have used reward expectancy and conditioning tasks to identify brain responses that correlate directly with rewards, reward-prediction-error signals (critic), and signals related to reward-dependent actions (actor). Many of these experiments have used reinforcement learning models as a way to understand the resulting brain responses, to choose design details of the experiment, or to locate brain responses associated with specific model components[30–34].

Human reward responses

Responses to rewarding stimuli have been observed consistently from the same set of subcortical regions in human brains, suggesting that neurons in these regions respond to a wide spectrum of triggers. In a series of elegant papers, Breiter and colleagues used fMRI to record brain responses to beautiful visual images[35] and drugs that induce euphoria (cocaine)[36]. The brain structures they identified included the orbitofrontal cortex (OFC), amygdala (Amyg), nucleus accumbens (NAc; part of the ventral striatum), sublenticular extended amygdala (SLEA; part of the basal forebrain), ventral tegmental area (VTA), and hypothalamus (Hyp). All these regions have topographically organized reciprocal connections with the VTA — one of the primary dopaminergic nuclei in the brainstem.

Particularly strong reward responses have been observed in the ventral striatum where numerous studies have shown that even abstract proxies for reward (money) cause activations that scale in proportion to reward amount or deviation from an expected payoff[37–39]. Similar results have been found by a variety of groups using both passive and active games with monetary payoffs[40–42]. A prominent activation response to monetary payoff was observed by Knutson and colleagues in the NAc and is shown in Fig. 3. The NAc, like the OFC and other parts of the prefrontal cortex (PFC), is densely innervated by dopaminergic fibres originating from neurons housed in the midbrain. Other work has shown that simply changing the predictability of a stimulus will activate the NAc and surrounding structures in the ventral parts of the striatum[30]. The picture emerging from this work is that responses in this region may reflect an encoding of rewards along a common valuation scale[43].

Human critic responses

One of the most important contributions of reinforcement learning theory has been to distinguish between the signalling of the reward itself, and the computation of the reward-prediction error. Using passive tasks with a juice reward, reward-prediction errors have been shown to activate structures in the ventral striatum[30,44]. Recently, two independent groups used passive learning paradigms to visualize reward-prediction-error signals in overlapping regions of the ventral putamen[32,33] (Fig. 4). The cingulate cortex is another area that has been associated with reinforcement learning signals that seem to be reward-prediction errors. The error-related negativity (ERN) is a scalp-recorded event-related potential (ERP), believed to originate from the anterior cingulate cortex, that is consistently observed about 100 msec following the commission of an error[45,46]. Similar potentials have been observed following negative feedback or unexpected losses in gambling tasks[47–49]. Holroyd and Coles have proposed that these potentials reflect a negative reward-prediction-error signal, and this idea has been tested under a variety of conditions[50–52]. Recently, fMRI evidence has suggested that a region of anterior cingulate cortex responds under many of the same conditions as the ERN: activity is affected by both errors and negative feedback[53].

Human actor responses

One implication of reinforcement theory for behaviour concerns the relationship between reward-prediction errors (critic signals) and action selection (actor signals). As discussed in Box 1, the critic signal can be used for reward learning and to adjust the future selection of reward-yielding actions. Success in the use of fMRI to detect reward-prediction-error signals inspired O'Doherty and colleagues to carry out a clever, but simple experiment designed to relate critic signals to action selection[34]. The experiment used a conditioning paradigm that was carried out in two modes. The first required an action to obtain a juice reward and the second did not. This experiment showed that activity in the dorsal striatum correlated with the prediction-error signal only when an action was needed to acquire the juice reward (Fig. 4c). There was no similar activity in this area when the juice was passively delivered. This finding is important because the dorsal striatum is involved in the selection and sequencing of actions.

Neuromodulation and cognitive control

Our consideration of reinforcement learning theory so far has focused on simple situations, involving the association of a stimulus with a reward, or with the selection of an action that leads to an immediate reward. In the real world, however, accrual of reward may require an extended sequence of actions. Furthermore, we have considered only a highly abstracted definition of the goal of the organism — the maximization of cumulative future rewards. However, many different forms of reward (and associated actions) may be

valued by an organism (for example, the procurement of nutrition, provision of safety, reproduction). This suggests that the construct of a goal needs to be refined to describe the variety of goal-directed behaviour in which humans engage. The guidance of behaviour in the service of internally represented goals or intentions, is often referred to as the capacity for cognitive control. Recent theories of cognitive control have elaborated on basic reinforcement learning mechanisms to develop models that specifically address the two challenges suggested above: (1) the need to learn and control sequences of actions required to achieve a goal; and (2) the need to represent the variety of goals that an organism may value. Here, we focus on the first of these challenges, but see refs 54 and 55 for a discussion of the latter.

Prefrontal goals

Pursuit of a goal (for example, going to the car, driving to the grocery store, or locating the refrigerated section to buy milk), can often require an extended sequence of actions. Theories of cognitive control consistently implicate the PFC as a site where representations of goals are actively maintained and used to select goal-directed behaviours[54]. The involvement of the PFC is motivated by three diverse classes of observations: (1) the PFC can support sustained activity in the face of distracting information[56,57]; (2) damage to the PFC produces deficits in goal-directed behaviour[58,59]; and (3) the PFC is selectively engaged by tasks that rely heavily on the active representation of goal information[60].

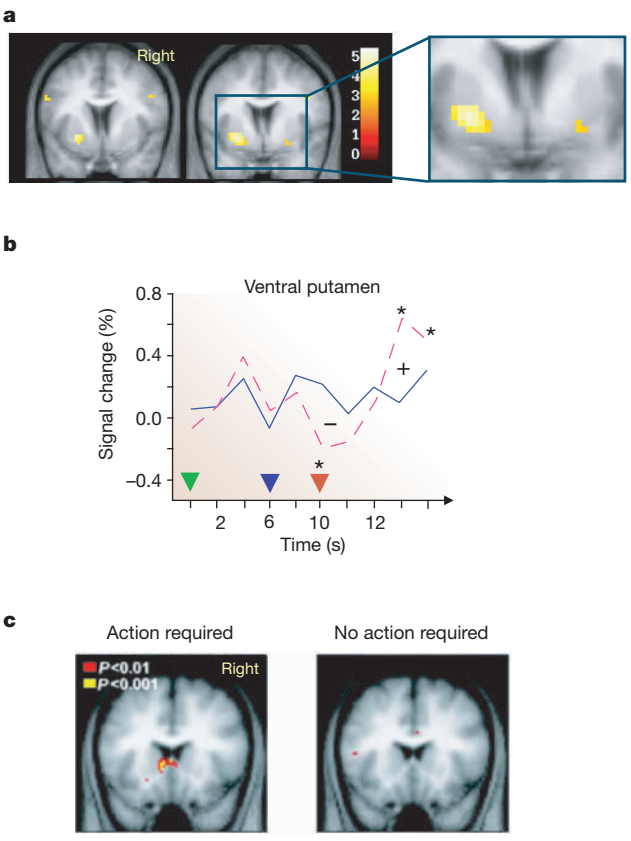

Figure 4 Detecting actor and critic signals in the human brain using fMRI. **a**, A simple conditioning task reveals a TD-like prediction-error response (critic signal) in the human brain. A cue is followed by the passive delivery of pleasant-tasting juice while subjects are scanned. The highlighted activation is located in the ventral part of the striatum (the putamen) — a region known to respond to a range of rewards. The activation represents the brain response that correlates with a continuous TD-like error signal. Adapted from ref. 30. **b**, A similar experimental design, but in this case a single prediction error of each polarity (positive and negative) can be seen in the ventral putamen during a surprising catch trial. Predictive sensory cue (green arrowhead); normal reward-delivery time (blue arrowhead); delayed reward time on catch trials (red arrowhead). Average BOLD (blood oxygenation level dependent) response in normal trials (solid line) and delay trials (dashed line). Adapted from ref. 32. **c**, Identification of actor response in dorsal striatum. A conditioning task is carried out in two modes requiring: (1) a button press (an action); and (2) no action at all. The dorsal striatum — a region involved in action selection — responds only during the mode where action is required and shows no response when action is not required. This is the first demonstration of an actor response detected in the human brain. Adapted from ref. 33.

Dopamine gating hypothesis

One problem with the simple hypothesis that the PFC actively maintains goal representations is that this does not indicate how or when this information should be updated. Failure to appropriately update goal representations will lead to perseverative behaviour, whereas failure to adequately maintain them will result in distractability. Indeed, disturbances of the PFC are known to be associated with distractability, perseveration, or both[61]. What is required is a mechanism that can signal when the goal representation should be updated. Recently, it has been proposed that dopaminergic signals from the VTA implement this mechanism, by controlling the 'gating' of afferent information into the PFC[55,62] (Fig. 5). According to this gating hypothesis, the PFC is resistant to the influence of afferent signals in the absence of phasic dopamine release, allowing it to preserve the currently maintained goal representation against impinging sources of interference. However, stimuli that signal the need to update the goal representation elicit a phasic dopamine response that 'opens the gate' and allows afferent signals to establish a new goal representation in the PFC.

Reinforcement learning and working memory

How does the dopamine system know which stimuli should elicit a gating signal and which should not? One plausible answer to this question comes directly from the reinforcement learning theory of dopamine function. A gating signal is required to update the PFC when a stimulus occurs in the environment which indicates that a more valuable goal can be achieved if behaviour is redirected towards that goal (for example, a light signalling that a reward can be acquired by going to some new location). In reinforcement learning terms, this corresponds to a positive reward-prediction error: the value of the current state is better than expected. According to the reinforcement learning theory of dopamine function, this is associated with a phasic burst in dopamine activity. In other words, reinforcement learning theory predicts that phasic dopamine responses will occur precisely when needed to produce a gating signal. Furthermore, insofar as the phasic dopamine response acts as a learning signal, it will strengthen the association of the current predictor, for example, the light, with the goal representation in the PFC. It will also strengthen the tendency of the light to elicit a phasic dopamine response when it recurs in the future. The learning here is analogous to the simple 'light-predicts-juice' experiments described earlier, except that now 'light predicts goal representation in the PFC', which in turn leads to the accrual of reward. This proposal shows how a prefrontal representation that plays a causal role in the acquisition of some later reward comes to be selected and reinforced.

Assuming that dopamine generates both learning and gating effects, the dopamine system provides a mechanism for learning which stimuli should elicit a gating signal to update goal representations in the PFC. Consistent with this hypothesis, the parameter used to implement the learning effects of dopamine in formal models of reinforcement learning[2,8,30,63] bears a remarkable similarity to the parameter used to implement gating effects in models of dopamine-based gating signals in the PFC[63]. Recent computational modelling work has demonstrated that implementing concurrent effects of dopamine phasic signals on reinforcement learning and gating allows a system to associate stimuli with the gating signals that predict reward, and so learn how to update representations appropriately in the PFC[62,64,65].

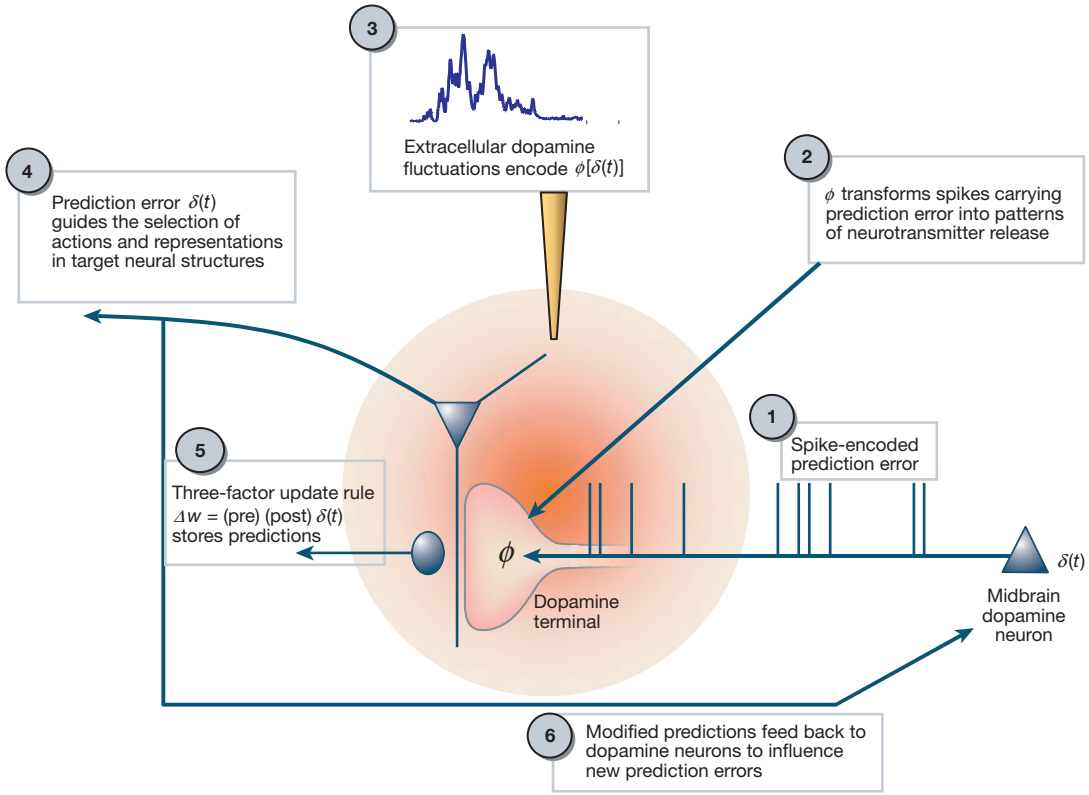

Figure 5 The flow and transformation of signals carried by the dopaminergic system. This system is now thought to be one part of a large, sophisticated neural system for valuation. (1) Dopamine neurons encode reward-prediction-error signals as modulations in their baseline firing rate; (2) transformation ϕ characterizes the way in which modulation of firing rate changes dopamine delivery (ϕ is known to be non-linear)[72]; (3) movement of dopamine through the extracellular space carries prediction-error information away from the synapse; (4) dopamine delivery to target structures controls a range of functions including the gating of working memory and the selection of specific actions; (5) any multiplicative learning rule that depends on the dopamine-encoded prediction error is able to store predictions, a vast improvement over simple storage of correlations familiar from hebbian learning; (6) changes in target structures act to adjust predictions, which are delivered back to dopamine neurons through long-range connections.

Recent work has begun to explore the hypothesis that the basal ganglia provide a mechanism for selective updating of goal representations within the PFC. This proposes that an important component of dopaminergic gating takes place in the basal ganglia, acting selectively on recurrent pathways that run from the PFC through the basal ganglia and back to the PFC. Computational models of the basal ganglia have shown how this system can learn tasks that require hierarchical updating of goal representations.

Neuromodulation and pathologies of cognitive control

Reinforcement learning theory provides a formal framework within which to explore quantitatively the effects that alterations in dopamine function may have on behaviour. We consider here two disorders in which it has long been recognized that dopamine plays a major role: drug addiction and schizophrenia.

Disturbances of dopamine in addiction

Perhaps the best understood pathology of dopamine excess is drug addiction, which is defined as compulsive drug use despite serious negative consequences. Once a pattern of compulsion is established, it often proves remarkably persistent. Even when addicted individuals have been drug-free for extended periods, drug-associated cues can readily lead to relapse. Addictive drugs such as cocaine, amphetamine and heroin all increase dopamine concentrations in the NAc and other forebrain structures by diverse mechanisms[20,66] and are highly reinforcing.

A new way to conceptualize the process of addiction is in the terms described above[21,67]. If dopamine plays a central role in both stimulus–reward learning and stimulus–action learning, and addictive drugs result in greater and longer-lasting synaptic dopamine concentrations than any natural reward, several predictions follow. Cues that predict drug availability would take on enormous incentive salience, by means of dopamine actions in the NAc and PFC, and complex drug-seeking behavioural repertoires would be powerfully consolidated by dopamine actions in the dorsal striatum[21]. In addition, dopamine effects in the PFC may impair the ability of the addicted person to suppress prepotent drug-seeking behaviour[17]. Given that certain risk-associated behaviour produces phasic dopamine release, and given the similarities between the development of drug addiction and pathologic gambling, it is interesting that early human neuroimaging results suggest that similar brain circuits may be involved[68].

Collectively, these results point to a hijacking of dopamine signals in PFC and limbic structures by addictive drugs. Because these drugs directly engage dopamine-mediated reinforcement learning signals, they generate a feedback loop that reinforces behaviour leading to drug consumption, establishing a vicious cycle of action and learning that explains the compulsive nature of drug addiction. The degree to which these drugs disrupt both phasic and tonic dopamine signals is not yet clear. However, the reinforcement learning models described above provide a framework for considering possible effects. For the learning effects, over-training with cues that predict drug delivery is a natural consequence of the role of phasic dopamine in learning. The PFC gating signal would also be unnaturally disrupted by selecting and over-learning grossly maladaptive prefrontal representations. These two effects would conspire to yield a representation of the world that is grossly biased towards drug-related cues. In addition,

repeated selection of maladaptive prefrontal representations would catastrophically rearrange the way in which normal functions were categorized within the PFC. In this framework, the addicted person's PFC can no longer even categorize decision problems correctly, much less regain control over the choices that their nervous systems deem valuable. The advantage now is that the reinforcement learning models provide a parameterized view of these problems and may well yield new directions in future work.

Disturbances of dopamine in schizophrenia

Disturbances of dopamine function are also known to have a central role in schizophrenia. This was first suggested by the discovery of the neuroleptic drugs that are effective in ameliorating the hallucinations and delusions associated with this illness. The clinical efficacy of these drugs correlates directly with their potency in blocking dopaminergic neurotransmission[69]. Conversely, dopamine agonists (for example, L-dopa and amphetamines) reproduce some of the same symptoms of schizophrenia. Taken together, these results led to the hypothesis that schizophrenia is associated with a hyper-dopaminergic state. However, almost half a century of research has failed to provide solid support for this simple idea. Although neuroleptics treat some of the more dramatic symptoms of schizophrenia, they fail to treat the persistent and equally debilitating symptoms of the disease, including cognitive disorganization and avolition.

The failure of the classic dopamine hypothesis is perhaps not surprising, given our lack of understanding of the role that dopamine has in system-level function. The development of the formal models of dopamine function discussed above, and its interaction with other brain systems, offers hope for a more sophisticated understanding of how dopamine disturbances produce the patterns of clinical psychopathology observed in schizophrenia. For example, along with evidence of dopamine disturbances, it has long been recognized that schizophrenia is associated with disturbances of frontal lobe function. This was originally suggested by comparing disturbances in executive function observed in schizophrenia (for example, distractability, and cognitive disorganization) with those observed in patients with frontal lobe damage. More recently, neuro-imaging studies have produced more direct evidence of deficits in frontal lobe function, and several investigators have begun to link these deficits with disturbances of dopamine function.

Specifically, schizophrenia may be associated with reduced dopamine activity in frontal cortex coupled with excess dopamine activity in subcortical structures, such as the striatum[70]. Early modelling work showed how a reduction of dopaminergic gain modulation in the PFC can simulate the behavioural deficits observed in patients with schizophrenia[71]. The learning and gating functions of dopamine reviewed here suggest ways in which this theory could be elaborated to include specific neuropharmacological findings.

Perspective

Despite our growing knowledge about some of the biological disturbances associated with schizophrenia, as yet there is no biological assay that can be used to diagnose this disease definitively. This reflects the deep limitations in our understanding of the relationship between biological disturbances and their clinical expression as perturbed mental or emotional function. We are entering a time where the formal synthesis of experimental data, both behavioural and physiological, will be needed to address the many open questions surrounding mental illness and behavioural decision-making. □

doi:10.1038/nature03015

1. Sutton, R. S., & Barto, A. G. *Reinforcement learning* (MIT, Cambridge, Massachusetts, 1998).
2. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16,** 1936–1947 (1996).
3. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275,** 1593–1599 (1997).
4. Friston, K. J., Tononi, G., Reeke, G. N., Sporns, O. & Edelman, G. M. Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience* **59,** 229–243 (1994).
5. Houk, J. C., Adams, J. L., & Barto, A. G. in *Models of Information Processing in the Basal Ganglia* (eds Houk, J. C. Davis, J. L. & Beiser, D. G.) Ch. 13, 249–270 (MIT, Cambridge, Massachusetts, 1995).
6. Skinner, B. F. Behaviorism at fifty. *Science* **140,** 951–958 (1963).
7. Sutton, R. S. Learning to predict by the methods of temporal difference. *Mach. Learn.* **3,** 9–44 (1988).
8. Doya, K. Metalearning and neuromodulation. *Neural Netw.* **15,** 495–506 (2002).
9. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience* Ch. 9, 331–358 (MIT, Cambridge, Massachusetts, 2001).
10. Rescorla R. A. & Wagner A. R. in *Classical Conditioning 2: Current Research and Theory* (eds Black, A. H. & Prokasy, W. F.) 64–69 (Appleton Century-Crofts, New York, 1972).
11. Bertsekas, D. P. & Tsitsiklis, J. N. in *Neuro-Dynamic Programming* (Athena Scientific, Belmont, Massachusetts, 1996).
12. Schultz, W. Apicella, P. & Ljungberg, T. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J. Neurosci.* **13,** 900–913 (1993).
13. Hollerman J. R. & Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neurosci.* **1,** 304–309 (1998).
14. Schultz, W. Predictive reward signal of dopamine neurons. *J. Neurophysiol.* **80,** 1–27 (1998).
15. Waelti, P., Dickinson, A. & Schultz, W. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* **412,** 43–48 (2001).
16. Bayer, H. M. & Glimcher, P. W. Subjective estimates of objective rewards: using economic discounting to link behavior and brain. *Soc. Neurosci. Abstr.* **28,** 358.6 (2002).
17. Berridge, K. C. & Robinson, T. E. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res. Rev.* **28,** 309–369 (1998).
18. Everitt, B. J. *et al.* Associative processes in addiction and reward: the role of amygdala-ventral striatal subsystems. *Ann. NY Acad. Sci.* **877,** 412–438 (1999).
19. Ikemoto, S. & Panksepp, J. The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. *Brain Res. Rev.* **31,** 6–41 (1999).
20. Di Chiara, G. & Imperato, A. Drugs abused by humans preferentially increase synaptic dopamine concentrations in the mesolimbic system of freely moving rats. *Proc. Natl Acad. Sci. USA* **85,** 5274–5278 (1988).
21. Berke, J. D. & Hyman, S.E. Addiction, dopamine, and the molecular mechanisms of memory. *Neuron* **25,** 515–532 (2000).
22. Ikemoto, S. & Panksepp, J. Dissociations between appetitive and consummatory responses by pharmacological manipulations of reward-relevant brain regions. *Behav. Neurosci.* **110,** 331–345 (1996).
23. Salamone, J. D. & Correa, M. Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. *Behav. Brain Res.* **137,** 3–25 (2002).
24. Redgrave, P., Prescott, T. J. & Gurney, K. Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci.* **22,** 146–151 (1999).
25. Egelman, D. M., Person, C., Montague & P. R. A computational role for dopamine delivery in human decision-making. *J. Cogn. Neurosci.* **10,** 623–630 (1998).
26. McClure, S. M., Daw, N. & Montague, P. R. A computational substrate for incentive salience. *Trends Neurosci.* **26,** 423–428 (2003).
27. Balleine, B. W. & Dickinson, A. The effect of lesions of the insular cortex on instrumental conditioning: evidence for a role in incentive memory. *Neurosci.* **20,** 8954–8964 (2000).
28. Berridge, K.C. in *The Psychology of Learning and Motivation: Advances in Research and Theory* Vol. 40 (ed. Medin, D. L.) 223–278 (Academic, San Diego, 2001).
29. Dayan, P. & Balleine, B. W. Reward, motivation and reinforcement learning. *Neuron* **36,** 285–298 (2002).
30. Berns, G. S., McClure, S. M., Pagnoni, G. & Montague, P. R. Predictability modulates human brain response to reward. *J. Neurosci.* **21,** 2793–2798 (2001).
31. O'Doherty, J. P., Deichmann, R., Critchley, H. D. & Dolan, R. J. Neural responses during anticipation of a primary taste reward. *Neuron* **33,** 815–826 (2002).
32. O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. Temporal difference models and reward related learning in the human brain. *Neuron* **38,** 329–337 (2003).
33. McClure, S. M., Berns, G. S., & Montague, P. R. Temporal prediction errors in a passive learning task activate human striatum. *Neuron* **38,** 339–346 (2003).
34. O'Doherty, J. P. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304,** 452–454 (2004).
35. Aharon, I. *et al.* Beautiful faces have variable reward value: fMRI and behavioral evidence. *Neuron* **32,** 537–551 (2001).
36. Breiter H. C. *et al.* Acute effects of cocaine on human brain activity and emotion. *Neuron* **19,** 591–611 (1997).
37. Breiter, H. C., Aharon, I., Kahneman, D., Dale, A. & Shizgal, P. Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* **30,** 619–639 (2001).
38. Knutson, B., Westdorp, A., Kaiser, E. & Hommer, D. fMRI visualization of brain activity during a monetary incentive delay task. *Neuroimage* **12,** 20–27 (2000).
39. Knutson, B., Adams, C. M., Fong, G. W. & Hommer, D. J. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neurosci.* **15,** 1–5 (2001).
40. Thut, G. *et al.* Activation of the human brain by monetary reward. *Neuroreport* **8,** 1225–1228 (1997).
41. Delgado, M. R., Nystrom, L. E., Fissel, C., Noll, D. C. & Fiez, J. A.Tracking the hemodynamic responses to reward and punishment in the striatum. *J. Neurophysiol.* **84,** 3072–3077 (2000).
42. Elliott, R., Friston, K. J. & Dolan, R. J. Dissociable neural responses in human reward systems. *J. Neurosci.* **20,** 6159–6165 (2000).
43. Montague, P. R. & Berns, G. S. Neural economics and the biological substrates of valuation. *Neuron* **36,** 265–284 (2002).
44. Pagnoni, G., Zink, C. F., Montague, P. R. & Berns, G. S. Activity in human ventral striatum locked to errors of reward prediction. *Nature Neurosci.* **5,** 97–98 (2002).
45. Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E. & Donchin, E. A neural system for error detection and compensation. *Psychol. Sci.* **4,** 385–390 (1993).
46. Falkenstein, M., Hohnsbein, J. & Hoormann, J. in *Perspectives of Event-Related Potentials Research* (eds Karmos, G. *et al.*) 287–296 (Elsevier Science, Amsterdam, 1994).
47. Gehring, W. J. & Willoughby, A. R. The medial frontal cortex and the rapid processing of monetary gains and losses. *Science* **295,** 2279–2282 (2002).
48. Ullsperger, M. & von Cramon, D. Y. Error monitoring using external feedback: specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging. *J. Neurosci.* **23,** 4308–4314 (2003).

49. Nieuwenhuis, S., Yeung, N., Holroyd, C. B., Schurger, A. & Cohen J. D. Sensitivity of electrophysiological activity from medial frontal cortex to utilitarian and performance feedback. *Cereb. Cort.* **14,** 741–747 (2004).
50. Holroyd, C. B. & Coles, M. G. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* **109,** 679–709 (2002).
51. Holroyd, C. B., Nieuwenhuis, S., Yeung, N. & Cohen, J. D. Errors in reward prediction are reflected in the event-related brain potential. *Neuroreport* **14,** 2481–2484 (2003).
52. Holroyd, C. B., Larsen, J. T. & Cohen, J. D. Context dependence of the event-related brain potential associated with reward and punishment. *Psychophysiol.* **41,** 245–253 (2004).
53. Holroyd C. B. *et al.* Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature Neurosci.* **7,** 497–498 (2004).
54. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. of Neurosci.* **24,** 167–202 (2001).
55. O'Reilly, R. C., Braver, T. S., & Cohen, J. D. in *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (eds Miyake, A. & Shah, P.) Ch. 11, 375–411 (Cambridge Univ. Press, New York, 1999).
56. Miller, E. K., Li, L. & Desimone, R. A neural mechanism for working and recognition memory in inferior temporal cortex. *Science* **254,** 1377–1379 (1991).
57. Miller, E. K., Erickson, C. A. & Desimone, R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* **16,** 5154–5167 (1996).
58. Duncan, J. Disorganization of behavior after frontal lobe damage. *Cog. Neuropsychol.* **3,** 271–290 (1986).
59. Shallice, T. in *From Neuropsychology to Mental Structure* (Cambridge Univ. Press, Cambridge, 1988).
60. Koechlin, E., Ody, C. & Kouneiher, F. The architecture of cognitive control in the human prefrontal cortex. *Science* **302,** 1181–1185 (2003).
61. Stuss, D. T. & Knight, R. T. *Principles of Frontal Lobe Function* (Oxford Univ. Press, New York, 2002).
62. Braver, T. S. & Cohen, J. D. in *Attention and Performance XVIII; Control of Cognitive Processes* (eds Monsell, S. & Driver, J.) 713–737 (MIT, Cambridge, Massachusetts, 2000).
63. Daw, N. D., Kakade, S. & Dayan, P. Opponent interactions between serotonin and dopamine. *Neural Netw.* **15,** 603–616 (2002).
64. O'Reilly, R. C., Noelle, D. C., Braver, T. S. & Cohen, J. D. Prefrontal cortex and dynamic categorization tasks: representational organization and neuromodulatory control. *Cereb. Cort.* **12,** 246–257 (2002).
65. Rougier, N. P. & O'Reilly, R. C. Learning representations in a gated prefrontal cortex model of dynamic task switching. *Trends Cogn. Sci.* **26,** 503–520 (2002).
66. Wise, R. A. & Bozarth, M. A. A psychomotor stimulant theory of addiction. *Psychol. Rev.* **94,** 469–492 (1987).
67. Hyman, S. E. & Malenka, R. C. Addiction and the brain: the neurobiology of compulsion and its persistence. *Nature Rev. Neurosci.* **2,** 695–703 (2001).
68. Potenza, M. N. *et al.* Gambling urges in pathological gambling: a functional magnetic resonance imaging study. *Arch. Gen. Psych.* **60,** 828–836 (2003).
69. Cohen, B. Dopamine receptors and antipsychotic drugs. *Mclean Hosp. J.* **6,** 95–115 (1981).
70. Weinberger, D. R. Implications of normal brain development for the pathogenesis of schizophrenia. *Arch. Gen. Psych.* **44,** 660–669 (1987).
71. Servan-Schreiber, D., Printz, H. & Cohen J. D. A network model of catecholamine effects: gain, signal-to-noise ratio and behavior. *Science* **249,** 892–895 (1990).
72. Montague, P. R. *et al.* Dynamic gain control of dopamine delivery in freely moving animals. *J. Neurosci.* **24,** 1754–1759 (2004).

Competing interests statement The authors declare that they have no competing financial interests.