

 Open access • Book Chapter • DOI:10.1016/B978-0-444-63425-2.00003-9

The role of learning-related dopamine signals in addiction vulnerability.

— [Source link](#) 

Quentin J. M. Huys, Philippe N. Tobler, Gregor Hasler, Shelly B. Flagel ...+1 more authors

Institutions: University of Zurich, University of Bern, University of Michigan,
Molecular and Behavioral Neuroscience Institute

Published on: 01 Jan 2014 - Progress in Brain Research (Elsevier)

Topics: Addiction vulnerability, Addiction and Incentive salience

Related papers:

- [A selective role for dopamine in stimulus–reward learning](#)
- [Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control](#)
- [Model-based influences on humans' choices and striatal prediction errors.](#)
- [A Neural Substrate of Prediction and Reward](#)
- [Reinforcement Learning: An Introduction](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/the-role-of-learning-related-dopamine-signals-in-addiction-3z6o5z2w4o>



**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2014

The role of learning-related dopamine signals in addiction vulnerability

Huys, Q J M ; Tobler, Philippe N ; Hasler, G ; Flagel, S B

Abstract: Psychostimulants such as methylphenidate (MPH) and antidepressants such as fluoxetine (FLX) are widely used in the treatment of various mental disorders or as cognitive enhancers. These medications are often combined, for example, to treat comorbid disorders. There is a considerable body of evidence from animal models indicating that individually these psychotropic medications can have detrimental effects on the brain and behavior, especially when given during sensitive periods of brain development. However, almost no studies investigate possible interactions between these drugs. This is surprising given that their combined neurochemical effects (enhanced dopamine and serotonin neurotransmission) mimic some effects of illicit drugs such as cocaine and amphetamine. Here, we summarize recent studies in juvenile rats on the molecular effects in the mid- and forebrain and associated behavioral changes, after such combination treatments. Our findings indicate that these combined MPH + FLX treatments can produce similar molecular changes as seen after cocaine exposure while inducing behavioral changes indicative of dysregulated mood and motivation, effects that often endure into adulthood.

DOI: <https://doi.org/10.1016/B978-0-444-63425-2.00003-9>

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-91359>

Journal Article

Accepted Version

Originally published at:

Huys, Q J M; Tobler, Philippe N; Hasler, G; Flagel, S B (2014). The role of learning-related dopamine signals in addiction vulnerability. *Progress in Neurobiology*, 211:31-77.

DOI: <https://doi.org/10.1016/B978-0-444-63425-2.00003-9>

The role of learning-related dopamine signals in addiction vulnerability

Quentin J.M. Huys^{1,2}, Philippe N. Tobler³, Gregor Hasler⁴, Shelly B. Flagel^{5,6}

1 Translational Neuromodeling Unit, Department of Biomedical Engineering, ETH Zürich and University of Zürich, Switzerland **2** Department of Psychiatry, Psychosomatics and Psychotherapy, Hospital of Psychiatry, University of Zürich, Switzerland **3** Laboratory for Social and Neural Systems Research, Department of Economics, ETH Zürich and University of Zürich, Switzerland **4** Department of Psychiatry, University of Bern, Switzerland **5** Department of Psychiatry, University of Michigan, Ann Arbor, Michigan, United States of America **6** Molecular and Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, United States of America

Corresponding Author

Quentin JM Huys
Translational Neuromodeling Unit, ETH Zürich
Wilfriedstrasse 6, 8032 Zürich, Switzerland
Email: qhuys@cantab.net

Word count

Abstract: 125
Total: 13475
Figures: 3; Tables: 2
References: 254

ABSTRACT

Dopaminergic signals play a mathematically precise role in reward-related learning, and variations in dopaminergic signalling have been implicated in vulnerability to addiction. Here, we provide a detailed overview of the relationship between theoretical, mathematical and experimental accounts of phasic dopamine signalling, with implications for the role of learning-related dopamine signalling in addiction and related disorders. We describe the theoretical and behavioural characteristics of model-free learning based on errors in the prediction of reward, including step-by-step explanations of the underlying equations. We then use recent insights from an animal model that highlights individual variation in learning during a Pavlovian conditioning paradigm to describe overlapping aspects of incentive salience attribution and model-free learning. We argue that this provides a computationally coherent account of some features of addiction.

CONTENTS

1 Background	2	4.5 Motivational shifts	14
1.1 Overview	3	4.6 Unlearning	15
2 Model-free and model-based learning from rewards	3	5 Individual variability	16
2.1 Model-based learning	4	5.1 Sign-tracking and goal-tracking	16
2.2 Model-free prediction error learning	4	5.2 Incentive salience accounts of the sign-tracking/goal-tracking variability	18
3 Phasic dopamine signals represent model-free prediction errors	7	5.3 Reinforcement learning accounts of the sign-tracking/goal-tracking variability	19
3.1 Causal role of (dopamine-mediated) prediction errors in learning	10	6 Addiction	21
3.2 Phasic dopamine signals in model-based learning	12	6.1 Phasic dopaminergic signals in addiction	22
4 Behavioural characteristics of model-free and model-based choices	12	6.2 Individual variability in addiction vulnerability	23
4.1 Outcome identity	12	6.3 Shifts towards model-free learning in addiction	24
4.2 Pavlovian approach and consummatory behaviours	13	6.4 Conclusions	25
4.3 Instrumental behaviour	14	6.5 Funding	26
4.4 Pavlovian-instrumental transfer	14	6.6 Acknowledgements	26
		References	26

1 BACKGROUND

Humans have used alcohol and various kinds of drugs of abuse for thousands of years. The early Egyptians consumed wine and narcotics, and the first documented use of marijuana in China dates back to 2737 B.C. However, the recognition of addiction as a problem occurred relatively recently and developed gradually in the 18th and 19th centuries (e.g. see Thomas de Quincey's "Confessions of an Opium Eater", 1821). The emergence of more potent formulations, better methods of delivery likely (Sulzer, 2011), and possibly expropriation of mechanisms aimed at internal regulation by drugs of abuse (Müller and Schumann, 2011) likely contributed to this development.

In today's societies, both legal and illicit drugs are readily available and most people experiment with potentially addictive drugs at some point in their lifetime. However, only a relatively small subset is vulnerable to developing addiction. Amongst those recently starting to use cocaine, for instance, about 5-6% are estimated to become cocaine abusers (O'Brien and Anthony, 2005). This subset nevertheless is of enormous impact, with addiction thought to affect at least 100 million individuals worldwide (Grant et al., 2004). Once affected, the consequences are severe, and relapse looms large. The most predictable outcome of a diagnosis of addiction is, unfortunately, not cure but a 90% chance of relapse (DeJong, 1994). Indeed, addiction represents a major public health concern with great consequences for physical and mental health, work and crime rates, resulting in a significant social and economic burden to society.

Historically, research into addiction has been multifaceted in terms of disease concepts and methods. Early on, addiction was considered primarily a social problem and was treated by legal measures and social institutions. The first criteria for a diagnosis of substance abuse and addiction were included in the third edition of the Diagnostic and Statistical Manual for the Classification of Mental Disorders (DSM-III) in 1980. Since then, the DSM has followed an "atheoretical" approach to provide reliable diagnoses for clinical practice, basing their diagnostic criteria for substance use disorders on clusters of clinical symptoms. Criteria include several aspects. One cluster of features centres on impairment of control over drug taking, which includes larger and longer drug use than originally intended; unsuccessful efforts to discontinue use; a great deal of time spent in substance use despite its consequences; and craving. Other clusters concentrate on the social impairments resulting from substance use, the risks drug-takers might expose themselves to as a direct consequence of drug effects, and also pharmacological criteria such as tolerance and withdrawal symptoms. With the exception of the type of drug and some pharmacological criteria, these symptom clusters have not been found to be directly associated with specific causes or pathogenetic processes. The newest version of DSM, DSM-5, states that an important characteristic of substance use disorders is an underlying change in brain circuitry that may persist beyond detoxification, particularly in individuals with severe disorders, without identifying what the specific underlying processes or "changes" might be. This chapter focuses on novel theoretical approaches and computational models from machine learning and decision theory in the hope that they might lend new scientific rigour to addiction research (Huys et al., 2011; Hasler, 2012). One of the beauties of addiction research is that the reinforcing effects of drugs of abuse and the development of drug dependence can be easily modelled in animals with high validity; and that theoretical frameworks are at a rather advanced stage of development.

Clinical, preclinical, epidemiological and theoretical findings suggest the importance of learning and neuroplasticity both in the pathogenesis of addiction disorders and in their cure. Specifically, the intake of a substance in larger amounts or over a longer period of time than originally intended and the persistent desire to cut down and regulate substance use may be considered as an expression of a learned involuntary habit; and result in reflexive thoughts and actions that contradict an individual's declared goals (Redish et al., 2008; Dayan and Niv, 2008; Graybiel, 2008). The understanding of learning processes has profited from computational modelling. This has supported the study of how individual variation in various forms of learning might underlie individual variation in the vulnerability to drug addiction. One insight gained from this work is that multiple learning processes occur in parallel and can, at least in part, be captured with so-called "model-free" and "model-based" learning theories. The so-called model-based learning system builds an

understanding of the world (Dayan and Niv, 2008; Balleine et al., 2009) in terms of what actions lead to what outcomes, akin to learning the rules of a game such as chess. In contrast, model-free learning systems allow adaptive behaviour in the absence of explanatory understanding. A shift from model-based towards model-free learning may be involved in the transition from occasional drug use to addiction. In the process, behaviour may become insensitive to changes in the subject's goals (Dayan and Niv, 2008). Indeed, maladaptive behaviours are characteristic of individuals with substance use disorders.

Dopamine is thought to play a pivotal role in these learning systems. Phasic dopaminergic signals appear to serve as teaching signals (Montague et al., 1996; Schultz et al., 1997) and be central to the attribution of incentive salience (Flagel et al., 2011b). The development of substance abuse and addiction likely involves the usurpation of such dopaminergic learning or incentive salience attribution signals (Volkow et al., 2009; Dayan, 2009; Flagel et al., 2011b). It has also been postulated that the attribution of incentive motivational value (i.e. incentive salience) to reward-associated cues contributes to the psychopathology of addiction. In the present chapter we review the role of dopamine in learning with a particular focus on its relevance to addiction. Emphasizing the important potential of theory-based and translational research approaches, we hope to illustrate how technological, theoretical and experimental approaches are bringing us closer to integrating the psychological and neurobiological processes underlying addiction vulnerability and relapse.

1.1 OVERVIEW

Part two of this paper reviews the standard reinforcement learning theory, focusing on so-called “model-free” and “model-based” decision-making (Sutton and Barto, 1998; Daw et al., 2005). We provide the mathematical foundation of these theories as a basis for subsequent interpretation of behavioural and neurobiological findings. The third part of the paper presents an overview over the evidence linking phasic dopaminergic responses to model-free learning. In the fourth part, we describe some important characteristics of these types of learning systems. Model-free learning is suggested to capture important aspects of both habits and incentive salience, while model-based learning is argued to relate to goal-directed valuation, be it instrumentally or in Pavlovian settings. Part five begins with a description of individual variability in a Pavlovian conditioning paradigm, whereby animals naturally segregate into those showing sign-tracking behaviour, or approach to a conditioned stimulus; versus goal-tracking behaviour or approach to the location of impending reward delivery. These findings are interpreted in light of two dominant theories: the reinforcement learning theory introduced in parts 2-4, and the incentive salience theory, presented in part 5. Finally, in the sixth part, we examine different paths to addiction arising from these data and models, focussing in particular on alterations to phasic signals reflecting terms from learning theory (reward prediction errors, i.e. the difference between expected and experienced reward), and a propensity towards model-free learning and incentive salience attribution.

2 MODEL-FREE AND MODEL-BASED LEARNING FROM REWARDS

Choosing behaviours that maximise rewards and minimize losses in the longer term is the central problem that reinforcement learning theory (RL) addresses. A difficulty in doing so is the appropriate balancing of short-term gains against long-term losses. Choices made now can have many different consequences tomorrow. The choice to enjoy another drink now may lead to social disinhibition and facilitate friendships or encounters, but it may also impair the ability to fulfil duties at work the next day, with more long-term negative impacts on the ability to maintain a social role. Patients with addiction have major difficulties striking this bargain (Kirby et al., 1999). RL theory provides one path to identifying adaptive decisions that take both long- and short-term consequences of choices into account. In particular, it addresses the problem that there are many possible futures that need to be considered and appropriately weighted by the probability of materializing. RL theory thus attempts to formalize solution to problems addicts saliently fail to solve and hence forms a framework for thinking about them.

There are at present two fundamentally different classes of neurally plausible approaches to solve the reinforcement learning problem: model-based and model-free learning. As we will detail below, model-based learning solves the reinforcement learning problem (i.e. the approach to maximising rewards and minimising losses in the longer term) by explicitly considering all future consequences of different actions. A typical example would be considering all possible sequences of moves in a game of chess. This is hampered by the huge computational costs this requires (Figure 1). Model-free learning solves the reinforcement learning problem in a more affordable manner, but this benefit comes at a large experiential cost: it suffers from the need for extensive, slow, sampling of the environment. Instead of considering all possible moves hypothetically, the consequences of the moves need to be experienced empirically by the model-free system.

2.1 MODEL-BASED LEARNING

Model-based decision-making involves building a model of the outcomes of actions and using this to infer the best sequence of actions. Consider a simple environment, in which only a few actions are available, each with three different outcomes, and each leading to another set of available actions. The task is to select the best sequence of two actions (Figure 1A). In its simplest incarnation, model-based decision-making corresponds to sequentially evaluating all possible action sequences and choosing the best. This demands a so-called **model** of the world, which in turn consists of two parts. First, a **transition matrix** \mathcal{T}^a describes the possible consequences of each action a . In Figure 1B,C examples are given of how transition matrices describe what actions lead to what outcomes with what probability. Second, it encompasses a **reward matrix** \mathcal{R} that describes the reinforcements for taking an action in a particular state. What is central to this representation is that the causal structure of the environment is captured in the set of all transition matrices $\mathcal{T} = \{\mathcal{T}^a\}$ for all actions, while the affective values are captured in \mathcal{R} and the two are represented separately. For a game, learning \mathcal{T} would consist of learning the rules, while learning \mathcal{R} would correspond to learning the aims of the game (in chess the capture of the opponent's king). Tree search would then require deriving the optimal play strategy from this information alone, notably without actually needing to experience playing the game (Shallice, 1982; Simon and Daw, 2011; Huys et al., 2012).

Learning then corresponds to changing the model of the world, i.e. changing either \mathcal{T} or \mathcal{R} . Learning \mathcal{T} can happen in the absence of any rewards (Gläscher et al., 2010). That animals are able to do this was shown very elegantly in the classical work of Tolman (1948): animals that were exposed to a maze (but without food rewards hidden in it) were later faster at learning a route to a food reward than those not preexposed to the maze (Bouton, 2006). However, the number of branches in a decision tree scales as w^d where w is the width of one level, and d the length of the action sequence. For a game such as chess the width is around 30, and the length of a game up to 40 moves long, rendering simple applications of this approach computationally suicidal. Nevertheless, for small problems, such as determining the fastest way to the nearest coffee shop from your office, it is feasible. Thus, sequential evaluation or tree search consists of acquiring a model of the world, and searching this to infer adaptive behaviours. It is resource intensive and limited to rather small decision problems, but it rapidly and efficiently reflects new information as long as the new information can efficiently be used to alter \mathcal{T} or \mathcal{R} .

2.2 MODEL-FREE PREDICTION ERROR LEARNING

The second approach to maximising reward relies on iterative updates via model-free prediction errors. Prediction errors are the difference between what one expects and what one gets. Casually put, imagine you order your favourite pizza at a restaurant (say with gorgonzola and pears) and instead are served with a different, less-preferred pizza (say with ham and pineapples). This would constitute a negative prediction error where the eventuality is worse than the anticipation. If, however, the waiter then apologized, brought you a pizza you liked as much as your preferred pizza and threw in two free beers you might experience a positive prediction error, with the outcome being better than your expectation. There would be no prediction error if you got just a pizza

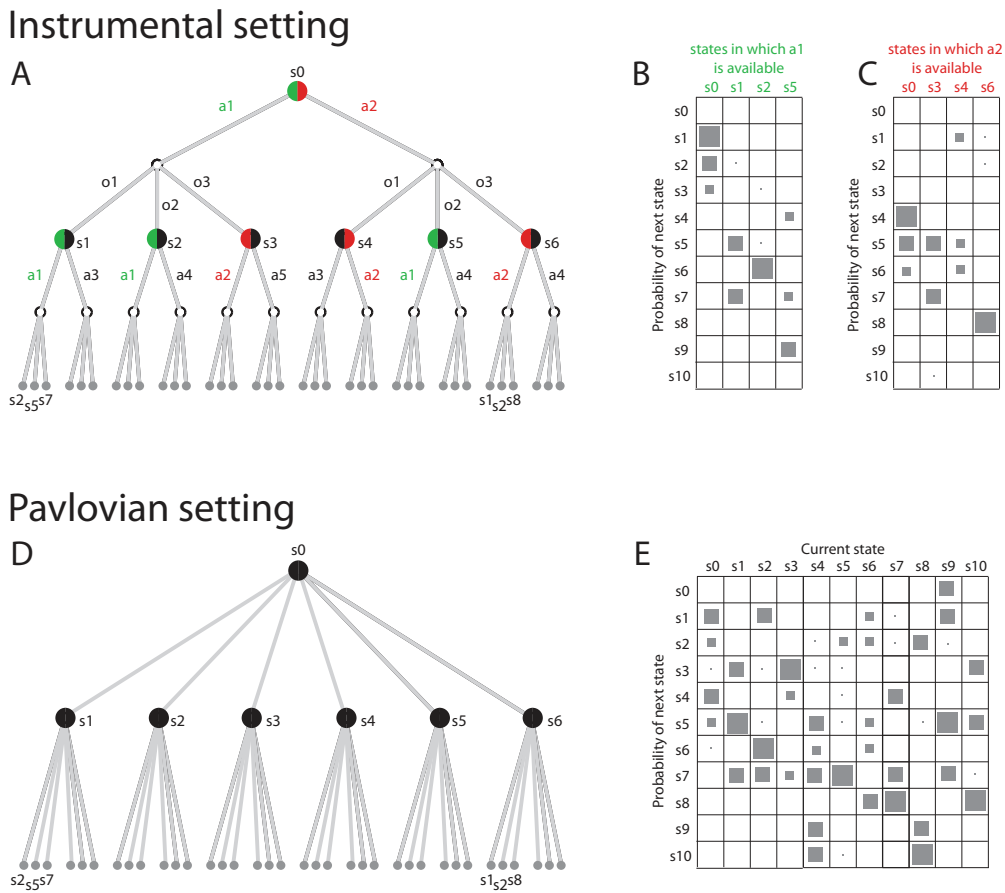


FIGURE 1: Model-based decision-making can be depicted as a decision-tree in both instrumental and Pavlovian settings. **A:** In an instrumental setting, model-based decision-making would consider all possible action sequences by evaluating all branches of the tree and determining the best one. The figure shows a specific instance, where the problem consists of first choosing between actions a_1 and a_2 , each of which has three possible outcomes, leading to three different states (s_1 - s_6). In each state, there is then a further choice between two actions, though different states have different available actions. Each of these actions in turn has three further possible outcomes, where the probability of each outcome depends on the state in which the action was taken. Actions are shown as solid circles, with green indicating that action a_1 is available, and red that action a_2 is available. Empty circles are outcome nodes. In order to choose the optimal sequence of actions, a goal-directed decision-making system has to consider all the options corresponding to all the branches in this decision-tree. In this simple problem, with a sequence of two choices, each leading to three possible outcomes, the tree has width $w = 6$, depth $d = 2$ and $w^d = 36$ branches. Thus, the difficulty of the problem is exponential in the length of the action sequence considered. **B,C:** Example transition matrices \mathcal{T} for actions a_1 and a_2 , respectively. Each column represents the probability distribution over next states when taking that action in a particular state. The larger the gray squares, the greater the probability. These transition matrices thus represent knowledge about action-outcome associations. There are similar matrices that describe when rewards are obtained. **D:** In an equivalent Pavlovian setting, model-based decisions would taken into account only state transitions. **E:** The model would now contain one single transition matrix \mathcal{T} describing the probability of going from one state to another, given a particular policy (behavioural strategy).

that you liked as much as the one you ordered (even if it's not the exact pizza you ordered). These prediction errors, and slightly more complex temporal versions of them, are used by the model-free system to acquire behaviour that is provably optimal in certain situations.

To properly understand the features of prediction error learning, it is worthwhile to consider it formally with a mathematical approach. To simplify the equations, we will consider using this approach to learn how much reward is associated with a stimulus or state s under Pavlovian conditions, but very similar equations describe learning for actions a or indeed state-action pairs (s, a) in instrumental conditioning. As explained above, optimal choices consider total future outcomes, not just immediate outcomes. This is formalized by considering the summed future outcomes $r_t + r_{t+1} + \dots$. Generally, however, the future is uncertain, and future rewards cannot simply be summed up. One must instead consider the average or expected total future reward $\mathbb{E}[r_t + r_{t+1} + \dots]$. This sum will be denoted as the value \mathcal{V} . Different states or situations are associated with different values, and hence we write the total expected future reward when in state s at time t as

$$\mathcal{V}(s_t) = \mathbb{E}[r_t + r_{t+1} + r_{t+2} \dots | s_t] \quad (1)$$

The right hand side of equation 1 can now be rewritten slightly differently as a sum of two terms. The first term is just the expected immediate reward $\mathbb{E}[r_t | s_t]$, while the second term contains the future rewards after the immediate reward, i.e. one and more time-steps into the future:

$$\mathcal{V}(s_t) = \mathbb{E}[r_t | s_t] + \mathbb{E}\left[\sum_{k=1}^{\infty} r_{t+k} \middle| s_t\right] \quad (2)$$

The key insight comes from equating the second term with the expected value of the next state s_{t+1} :

$$\mathcal{V}(s_t) = \mathbb{E}[r_t | s_t] + \mathbb{E}[\mathcal{V}(s_{t+1}) | s_t] \quad (3)$$

where the second expectation implies a weighting by (and sum over) the transition probability $P(s_{t+1} | s_t)$ of going from state s_t to another state s_{t+1} . This equation is key, as it tells us how the total future expected reward from state s_t (which previously we had to evaluate a large tree to obtain) is related to the total future reward from its successor states s_{t+1} : the difference should be exactly the expected immediate reward in state s_t . This equation, which is one form of the Bellman equation (Bellman, 1957; Sutton and Barto, 1998), thus provides a set of consistency checks between values of different states. It can now be used to learn by bootstrapping. Assume we have an incorrect value $\hat{\mathcal{V}}$. That means that equation 3 does not hold:

$$\mathcal{V}(s_t) \neq \mathbb{E}[r_t | s_t] + \mathbb{E}[\mathcal{V}(s_{t+1}) | s_t] \quad (4)$$

and that there is a difference Δ between the two sides:

$$\Delta = \mathbb{E}[r_t | s_t] + \mathbb{E}[\mathcal{V}(s_{t+1}) | s_t] - \mathcal{V}(s_t) \quad (5)$$

These equations involve expectations $\mathbb{E}[\cdot]$. The next insight, fundamental to RL techniques, is that this difference can be estimated by iteratively, over trial and error, averaging over actual experiences in the environment. Rather than computing one difference Δ , this is replaced by samples of the difference, called 'prediction errors' δ , where the $\delta \neq 0$ unless the values are correct (e.g. you receive the pizza you ordered). Learning occurs by slowly adding up these prediction errors δ over different visits to each state. Let t index time, with reward r_t experienced in state s_t followed by a transition to state s_{t+1} . Let $\hat{\mathcal{V}}_t(s_t)$ be the estimate of the value of that state s before the t 'th visit. Then equation 5 can be approximated by:

$$\delta_t = r_t + \hat{\mathcal{V}}_t(s_{t+1}) - \hat{\mathcal{V}}_t(s_t) \quad (6)$$

$$\hat{\mathcal{V}}_{t+1}(s_t) \leftarrow \hat{\mathcal{V}}_t(s_t) + \alpha \delta_t \quad (7)$$

If α is a small (< 1), but positive constant, then on consecutive visits the value of state s is always updated a little towards the value it should have, which is the true sum of the immediate reward

	Model-free	Model-based
Pavlovian (state) values	$\mathcal{V}^{\text{MF}}(s)$	$\mathcal{V}^{\text{MB}}(s)$
Instrumental (state-action) values	$\mathcal{Q}^{\text{MF}}(s, a)$	$\mathcal{Q}^{\text{MB}}(s, a)$

TABLE 1: Types of values. There are both Pavlovian state and instrumental state-action values, and both of these can be either model-free (cached) or model-based.

plus the value of future states. By doing this iteratively to a small extent on each visit i , equation 7 implements a running average.

We emphasize again that the prediction error measures the inconsistency of the current estimates $\hat{\mathcal{V}}(s_{t+1})$ and $\hat{\mathcal{V}}(s_t)$ with respect to the actually obtained reward r_t . Temporal difference (TD) prediction error learning implies measuring the inconsistency between how expectations change over time (the difference between the terms $\mathcal{V}(s)$ and $\mathcal{V}(s')$) and obtained rewards, and summing this up over many repeated experiences. The computations needing to be performed (equations 6-7) are now trivial. The work of evaluating the tree of future possibilities (as in model-based decision-making) has been shifted to experience rather than simulation based on rules. Hence, model-free prediction error learning trades computational cost for experiential cost.

A few further points about model-based and model-free learning deserve brief mention:

Knowledge Unlike the model-based tree search, the model-free equation 7 does not require a model of the world in terms of knowledge about action-outcome contingencies. Specifically, in order to learn, neither the transition matrix \mathcal{T} , nor the reward matrix \mathcal{R} have to be known - there only has to be the possibility of experiencing them. This corresponds merely to acting in the world and observing consequences and rewards.

State vs state-action values The model-free equations were written in terms of state values $\mathcal{V}(s)$, but could, with a few alterations, have been written in terms of state-action values, which are traditionally denoted by $\mathcal{Q}(s, a)$. Unlike state values $\mathcal{V}(s)$, these directly estimate how valuable a particular action is in a particular state. Just like model-free prediction-error based learning, model-based tree search can also be used to yield both state or state-action values. Figure 1D,E shows how the decision-tree in panel A can be formulated in terms of states only. That is, it is possible to use a model (\mathcal{T} and \mathcal{R}) to evaluate the expected reward of performing an action in a state, or the expected reward of being in a state (collapsing over possible actions).

Instrumental vs Pavlovian The model-free / model-based distinction is independent of the instrumental / Pavlovian distinction (Figure 1). In instrumental learning, subjects are reinforced for a stimulus-response combination, which is modelled using state-action values $\mathcal{Q}(s, a)$. In Pavlovian conditioning experiments, stimuli are predictive of reward irrespective of the actions emitted by the subjects. These stimulus-bound expectations are modelled using state values $\mathcal{V}(s)$. Clearly, the latter begs the question of how and why stimulus values elicit actions at all, and we will return to this below. However, we emphasize both model-based and model-free approaches can, in principle, be applied to either instrumental or Pavlovian scenarios. In other words, there can be both cached, model-free Pavlovian values $\mathcal{V}^{\text{MF}}(s)$ and instrumental values $\mathcal{Q}^{\text{MF}}(s, a)$ and model-based Pavlovian values $\mathcal{V}^{\text{MB}}(s)$ and instrumental values $\mathcal{Q}^{\text{MB}}(s, a)$.

3 PHASIC DOPAMINE SIGNALS REPRESENT MODEL-FREE PREDICTION ERRORS

The neural bases of model-based learning are not very clear, with only few direct measurements of tree search available (Johnson and Redish, 2007; van der Meer and Redish, 2009; Pfeiffer and Foster, 2013). However, the neural representation of prediction-error signals as required for model-free learning have been examined in exacting detail (Montague et al., 1996; Schultz et al., 1997),

and we turn to this evidence next. It focuses on the dopamine neurons of the ventral tegmental area (VTA) and, in a nutshell, suggests that dopamine neurons code some form of the δ term described above.

Dopaminergic involvement in reward learning has been studied with recordings of the electrical activity of single neurons, voltammetry (Day et al., 2007) and neuroimaging in rodents, macaques and humans. In now classical experiments (for reviews, e.g. Daw and Tobler 2013; Glimcher 2011; Schultz 1998, 2013), dopamine neurons were found to respond with a burst of action potentials (duration and latency of roughly 100 ms) to rewards such as small pieces of food hidden in a box or to drops of cordial delivered through a spout. While rewards typically activate dopamine neurons, punishments inhibit them (Fiorillo et al. 2013a,b; Fiorillo 2013; for review, e.g. Ilango et al. 2012, though see also Brischoux et al. 2009). When a reward (a unconditioned stimulus US) is repeatedly and consistently signalled by a sound or a visual stimulus (i.e. a conditioned stimulus, CS), the phasic activation no longer occurs at the time when the reward is received, but instead transfers to the onset time of the CS. This parallels how the prediction errors δ in the model-free account would behave: initially, the reward is unexpected, and hence leads to a positive prediction error. After learning occurs, the presentation of the CS at unpredictable times leads to positive prediction errors, but the reward itself (which is predicted by the CS and hence no longer surprising) fails to elicit a dopamine response. The response transfer from CS to US parallels the development of conditioned behaviour (e.g. conditioned licking with liquid reward) in response to presentation of the CS during learning.

Multiple features of dopamine firing align closely with model-free accounts. The phasic activation in response to CSs is independent of the sensory modality of the conditioned stimuli and increases with predicted reward magnitude (Bayer and Glimcher, 2005; Roesch et al., 2007; Tobler et al., 2005) and probability (Enomoto et al., 2011; Fiorillo et al., 2003; Morris et al., 2006; Nakahara et al., 2004; Satoh et al., 2003). This is again in line with the theoretical formulation as the expected value increases with the size of the reward and its probability. Furthermore, the longer the delay between the CS and the reward, the weaker the response (Fiorillo et al., 2008; Kobayashi and Schultz, 2008; Roesch et al., 2007), reflecting temporal discounting of future rewards. Finally, if a reward-predicting stimulus is itself preceded by another, earlier, stimulus, then the phasic activation of dopamine neurons transfers back to this earlier stimulus (Schultz et al., 1993), which is again captured by the above theoretical account (Montague et al., 1996) of model-free learning.

The relation to model-free learning is further illustrated by the finding that dopamine neurons not responding to reward predicted by conditioned stimuli nevertheless respond when reward occurs at unpredicted time points, e.g. outside the task or earlier than predicted (Hollerman and Schultz, 1998). Both of these situations constitute positive prediction errors and would be captured by a $\delta > 0$. Moreover, when reward is predicted but fails to occur (e.g. because it is withheld by the experimenter or because of an error of the animal), there is a negative error in the prediction of reward ($\delta < 0$). Dopamine neurons duly show a phasic depression in activity (Schultz et al., 1997; Tobler et al., 2003) and the duration of depressions increases with the size of the negative prediction error (Bayer et al., 2007; Mileykovskiy and Morales, 2011). Taken together, dopamine neurons seem to emit a model-free prediction error signal δ such that they are phasically more active than baseline when things are better than predicted (positive prediction error), less active than baseline when things are worse than predicted (negative prediction error) and show no change in activity when things are as good as predicted (no prediction error). In other words, the firing of dopamine neurons is well described by formal model-free approaches to reinforcement learning (equations 6- 7), suggesting that the dopaminergic prediction error signals not only an error in the prediction, but that it can also be used as a signal indicating precisely how much and in what direction expectations needs to be changed – a teaching signal (Sutton and Barto, 1998).

The activation elicited by the earliest reward-predicting stimulus can also be interpreted in terms of prediction error coding because the sudden occurrence of a reward-predicting stimulus constitutes a positive prediction error with respect to the preceding intertrial interval, during which no reward was predicted. In most experiments, the probability of reward at each moment in time is low due to relatively long and variable intertrial intervals. Reward-predicting stimuli induce positive prediction errors relative to that low background probability. Thus, dopamine neurons appear to

code errors in the prediction of reward at each moment in time as captured by equation 6. For instance, when a stimulus predicting reward at 25% is followed by either a stimulus predicting reward at 100% (positive prediction error) or another stimulus predicting 0% (negative prediction error), the second stimulus activates or depresses dopamine neurons, respectively (Takikawa et al., 2004). This finding further reinforces the notion that stimulus-induced activation of dopamine neurons strongly covary with prediction errors.

Many studies have confirmed, quantified, and extended reward prediction error coding by dopamine neurons, even in humans (Zaghloul et al., 2009). The dopamine neurons of monkeys that have not learned to predict reward show continued positive and negative prediction errors at the time of reward or reward omission, respectively. By contrast, the dopamine neurons of monkeys that have learned to predict reward well show conditioned stimulus responses indicative of learning in an asymmetrically rewarded saccade task (Kawagoe et al., 2004). In behavioral situations with contingencies changing about every 100 trials, dopamine neurons code the difference between current reward and reward history weighted by the last six to seven trials (Bayer et al., 2007). The occurrence of reward or reward prediction (positive prediction error) or their omission (negative prediction error) activates or depresses dopamine neurons in an inverse monotonic function of probability, such that the more unpredicted the event the stronger the response (de Lafuente and Romo, 2011; Enomoto et al., 2011; Fiorillo et al., 2003; Matsumoto and Hikosaka, 2009; Morris et al., 2006; Nakahara et al., 2004; Nomoto et al., 2010; Oyama et al., 2010; Satoh et al., 2003).

Enomoto et al. (2011) attempted to directly address whether the phasic dopamine response reflect the total future reward, as opposed to just the immediate reward. Monkeys first had to identify the currently reinforced target out of three possible targets by trial and error. They then received two or three further rewards for returning to that target. Equation 4 suggests that the predicted sum of future reward increases and decreases again as the monkeys progress through these exploration and exploitation trials. The suggestion is based on the expected value over the course of the trials and on the notion that later rewards are less valuable than sooner rewards. Both conditioned licking and phasic dopamine responses to the start cue of a trial closely follow the pattern suggested by the notion that they reflect time-resolved prediction errors not only about immediate rewards but, critically, the sum of immediate and future rewards, just as suggested by equation 5. These data demonstrate that dopamine neurons compute the prediction-error term with respect to a quantitative and time-resolved expected total future reward term $\mathbb{E}[\mathcal{V}(s_{t+1}|s_t)]$.

Enomoto et al. (2011) examined Pavlovian values in the setting of an instrumental task (c.f. Guitart-Masip et al. 2012). It is also possible to examine whether the phasic responses depend on what action was chosen, as should be the case in model-free instrumental acquisition of state-action $Q^{\text{MF}}(s, a)$ values. Indeed, dopamine neurons do show such a sensitivity (Morris et al., 2006; Roesch et al., 2007), and thus appear to be able to emit model-free prediction errors both when learning about stimulus values $\mathcal{V}^{\text{MF}}(s)$ as in Pavlovian settings, and when learning about stimulus-action values $Q^{\text{MF}}(s, a)$ as in instrumental settings.

Cyclic voltammetry has shown that dopamine release in the striatum, the main target region of dopamine neurons, follows many of the same features as the prediction error signals of dopamine neurons themselves (Day et al., 2007). In humans, functional MRI (fMRI) studies and even *in vivo* surgical recordings (Zaghloul et al., 2009) have reported correlates of prediction-error signals in the striatum that resemble those of dopamine neurons recorded in animals, including phasic (event-related) positive and negative prediction error responses (McClure et al., 2003a; O'Doherty et al., 2003; D'Ardenne et al., 2008) that reflect probability (e.g. Abler et al. 2006; Burke et al. 2010; Spicer et al. 2007; Tobler et al. 2007) and more specific predictions of formal learning theories (Daw et al., 2011; Kahnt et al., 2012; Rutledge et al., 2010; Tobler et al., 2007). However, it is worth keeping in mind that the haemodynamic response measured with neuroimaging is non-specific rather than a one-to-one reflection of a particular neural event such as dopamine release (see also Düzel et al. 2009), which could explain why some fMRI studies have suggested positive coding of losses (Seymour et al. 2004; although see also Tom et al. 2007) and a dominance of action over value (Guitart-Masip et al., 2012).

Blocking	Stage 1	Stage 2	Test
	A → reward	AX → reward	X?
	B → no reward	BY → reward	Y?
Optogenetic unblocking	Stage 1	Stage 2	Test
	A → reward	AX → reward + DA stimulation	X?
Transreinforcer blocking	Stage 1	Stage 2	Test
	A → shock	AX → shock	X?
	A → reward omission	AX → shock	X?
Identity unblocking	Stage 1	Stage 2	Test
	A → 3 units reward 1	AX → 3 units reward 2	X?

TABLE 2: Blocking designs. Learning to a stimulus, e.g. X is “blocked” by the presence of another stimulus A that already predicts the outcome. Stimuli are denoted by letters. The original blocking experiment (Kamin, 1969) used an aversive between-subjects design; by contrast, the experiment described in the text and depicted in abbreviated form here (Waelti et al., 2001) used an appetitive within-subject design where the test consists of a comparison between Y and X (see also Figure 2A); The optogenetic unblocking experiment of Steinberg et al. (2013) used a between-subject design. Here the test consisted of a comparison in the conditioned behaviour in response to presentation of X in three groups. In one group of rats the dopamine neurons were stimulated at the time of the reward in AX trials, while in the other groups the stimulation occurred at other times or not at all. In transreinforcer blocking (Dickinson and Dearing, 1979) and identity unblocking (McDannald et al., 2011), the reinforcer is changed at the AX compound stage. The test here consists of a comparison of behaviour in response to X after this change versus when no change has occurred (i.e. standard blocking).

3.1 CAUSAL ROLE OF (DOPAMINE-MEDIATED) PREDICTION ERRORS IN LEARNING

So far, we have argued that prediction errors play a role in model-free learning and that dopamine neurons emit a signal that closely resembles this formal prediction error. However, this falls short of showing that these prediction errors are indeed necessary for and causally involved in learning *in vivo*. One possibility of testing whether prediction errors are important for learning is to set up a behavioral situation in which two different stimuli are equally often paired with reward but only one of them is followed by a prediction error. This is exactly what the so-called “blocking paradigm” achieves (Kamin, 1969). In this paradigm, one new stimulus (denoted by the letter ‘X’ in the top row of Table 2) is added to a previously learned stimulus (‘A’) whereas another new stimulus (‘Y’) is added to a neutral stimulus (‘B’). Both compounds are followed by reward. After the compound with the pretrained stimulus (‘AX’) the reward occurs just as predicted by the pretrained stimulus (no prediction error) whereas after the compound with the neutral stimulus (‘BY’) the reward is unpredicted (positive prediction error). If prediction errors are important for learning, there should be no learning about the new stimulus ‘X’ in the former case but there should be learning about the new stimulus ‘Y’ in the latter case. Figure 2A shows that, in agreement with these predictions, monkeys show considerably less conditioned licking to stimuli that were not followed by a prediction error than to control stimuli that were followed by a reward prediction error (Waelti et al., 2001). Dopamine neurons show the same pattern (Figure 2A, bottom): they respond to stimuli that were followed by a prediction error but not to those that were not (Waelti et al., 2001). Thus, prediction errors are required for both stimulus-induced phasic activity of dopamine neurons and learning.

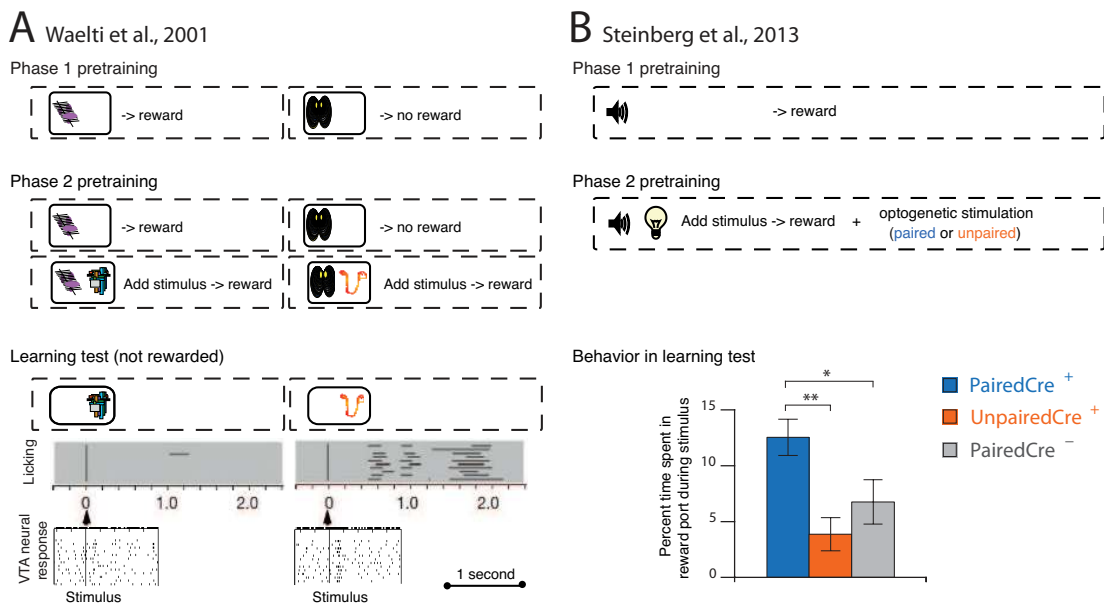


FIGURE 2: Dopamine neurons show blocking that parallels behavioral blocking. Learning is reinstated and blocking prevented when dopamine neurons are stimulated at the time of a predicted reward. **A:** Schematic of blocking task used with single neuron recordings from dopaminergic neurons in the VTA (within-subject design). In a first pretraining phase, a stimulus is paired with a drop of liquid reward (top left) whereas a control stimulus is not (top right). Accordingly, the animal forms an association between the left stimulus and reward but not between the right stimulus and reward. In a second pretraining phase, additional stimuli are occasionally presented together with the stimuli previously learned in the first pretraining phase. In this phase, both compounds are followed by reward. The reward elicits a prediction error in the control compound on the right but not in the experimental compound on the left. This is because the added stimulus is followed by unpredicted reward in the control but not in the experimental case. Because there is no prediction error, learning to the added stimulus on the left does not occur. In a third phase, the added stimuli were occasionally tested on their own (interspersed with the four trial types used during the pretraining phases in order to maintain learning). The blocked stimulus (left) and its control (right) are both followed by no reward and the behavior (conditioned licking, top) as well as the responses of a single dopamine neuron at the time of the stimulus (bottom) are shown. Control but not blocked stimuli elicit conditioned licking and phasic dopamine activations. Adapted with permission from Waelti et al. (2001). Note that haemodynamic responses in the striatum show a very similar response pattern (Tobler et al., 2006). **B:** Schematic of blocking task used with optogenetic stimulation (between-subject design). Pretraining phases proceeded similarly to the recording study, except that the nature of stimuli differed and in the second pretraining phase there were no reminders from the first pretraining phase. During the second phase, two groups received active stimulation of dopamine neurons concurrently with reward (PairedCre⁺) or during the intertrial interval (UnpairedCre⁺). A third group received inactive stimulation at the time of the reward (PairedCre⁻). The data are shown in the bar plot at the bottom as time spent in the reward port during stimulus presentation. The group with active stimulation at the time of the reward showed more Pavlovian approach behaviour than the other two groups, presumably due to the additional prediction error signal elicited by optogenetically induced phasic dopamine activity. Adapted with permission from Steinberg et al. (2013).

What remains is the question whether prediction error-like phasic dopamine responses are causally involved in reward learning? Recent evidence suggests they are. In an optogenetic variant of the blocking paradigm just described, dopamine neurons of rats were artificially activated at the time of the reward already predicted by the pretrained stimulus (Figure 2B, bottom; i.e. stimulation occurred in AX trials, c.f. second row of Table 2). If the prediction error hypothesis of dopamine firing is correct, this should induce an artificial prediction error at a time when no prediction error would have occurred naturally. As a result, this prediction error should lead to learning about the stimulus added to the pretrained stimulus. Indeed, rats in which this kind of stimulation was active showed stronger conditioned responding to the added cue on the first test trial than rats in which active stimulation was delivered during the intertrial interval rather than the time of reward or rats in which the appropriately timed stimulation was not active (Steinberg et al., 2013). Moreover, stimulation of dopamine neurons at the usual time of reward slowed behavioural extinction (Steinberg et al., 2013). Thus the stimulation counteracted the negative prediction error induced by the absence of expected reward and thereby conditioned behaviour was sustained. These findings support and extend previous optogenetic studies that implicated dopamine in learning by showing that dopamine neurons code reward prediction errors (Cohen et al., 2012), and that their activation is sufficient to reinforce intracranial self-stimulation (Kim et al., 2012; Rossi et al., 2013; Witten et al., 2011) and leads to conditioned place preference (Tsai et al., 2009) whereas inhibiting them causes avoidance learning (Tan et al., 2012).

3.2 PHASIC DOPAMINE SIGNALS IN MODEL-BASED LEARNING

The data discussed up to this point are in line with dopamine coding model-free, experiential prediction-errors. However, to some degree, dopamine responses incorporate information not available in current experiences into their prediction error responses. Consider a task in which the values of two stimuli are anticorrelated such that when one reverses from being rewarded to being unrewarded, the other automatically does the opposite (Bromberg-Martin et al., 2010). On the very first trial after realizing that the value of one stimulus has changed, a monkey can infer that the value of the other stimulus has also changed without having to experience the outcome of that stimulus (though note that this depends on representing the two stimuli separately). Both behaviour and dopamine neurons process inferred outcome values, although the impact of experienced value on both is more pronounced. In particular, dopamine neurons respond more strongly to a stimulus that is inferred to be valuable than to a stimulus that is inferred to be non-valuable. In a different task (Nakahara et al., 2004) as the number of unrewarded trials increases, the probability of reward increases. Instead of showing extinction, monkeys learn the structure of such a task and dopamine neurons track the probability of reward. These findings are consistent with dopamine neurons also playing some role in forms of model-based learning. We will return to this possibility below in the context of goal-tracking behaviour.

4 BEHAVIOURAL CHARACTERISTICS OF MODEL-FREE AND MODEL-BASED CHOICES

Above we have seen that phasic dopamine signals covary with a temporal difference prediction error. Henceforth, we will consider these signals as model-free. Model-free learning evaluates the total future reward by summing up the prediction errors over time into either $\mathcal{V}^{\text{MF}}(s)$ or $Q^{\text{MF}}(s, a)$ values. We briefly review several domains in which this has qualitative behavioural consequences that distinguish model-free from model-based choices.

4.1 OUTCOME IDENTITY

Model-free values $\mathcal{V}^{\text{MF}}(s)$ and $Q^{\text{MF}}(s, a)$ are nothing but the sum of past prediction errors. The error does not contain any information other than the discrepancy in the amount of reward obtained. Thus, $\mathcal{V}^{\text{MF}}(s)$ and $Q^{\text{MF}}(s, a)$ values arising from model-free learning do not contain any

other information, such as the identity of the reward. Model-free learning should thus be sensitive only to the size or valence of a reward, but not to its identity. This distinguishes it from the model-based system. In an aversive version of the blocking experiment (Table 2, top row; Kamin 1969), a stimulus A is first trained to predict shock. When a second stimulus, X, is added and the compound followed by shock, the ability of stimulus X to predict shock is reduced, even though it was paired with the shock, too. This provides behavioural evidence for the importance of prediction errors in learning. As discussed above, more direct evidence has recently been obtained by manipulating phasic dopamine signals (Steinberg et al., 2013) in an appetitive blocking design (Table 2, second row). In a variant of the blocking paradigm (transreinforcer blocking; Table 2, third row; Dickinson and Dearing 1979; Ganesan and Pearce 1988), the identity of the reinforcer is changed in the compound phase, e.g. from reward omission to shock presence. Strikingly, when A predicts the absence of reward, learning of the association between X and shock is blocked. This strongly suggests that 'reward' and 'punishment' are motivational opponents on a linear scale, and that in at least some types of learning the only aspect of the nature of the affective outcome (food reward or shock punishment) that is relevant is its value on that linear scale, and that other features are abstracted away.

However, animals are not entirely insensitive to the nature of Pavlovian outcomes and this can be revealed in other blocking experiments. In identity unblocking (Table 2, bottom row), the equivalence of two reward quantities (e.g. 1 pellet and 3 sucrose drops) is first assessed. A first CS is then conditioned to predict the first reward identity. Then, an identity shift occurs: the compound predicts the new reward, which was measured to be of equal value. Thus, there is no value prediction error (equation 7), yet animals are sensitive to such shifts (Seligman, 1970; Bouton, 2006; McDannald et al., 2011; Takahashi et al., 2011; Jones et al., 2012), showing that they do represent and learn more features about the outcome than the scalar measure of how rewarding it is. Thus, while transreinforcer blocking (and value blocking more generally) supports model-free processes, identity unblocking is evidence for model-based processes in Pavlovian conditioning (McDannald et al., 2011).

4.2 PAVLOVIAN APPROACH AND CONSUMMATORY BEHAVIOURS

Model-free Pavlovian state values $v^{\text{MF}}(s)$ do not contain explicit information about particular actions. They can nevertheless drive some simple behaviours, particularly when there is some distance between the organism and a positively valued stimulus. Unlike more complex behaviours evolved by Pavlovian stimuli, Pavlovian approach behaviours primarily involve locomotion to bring the organism closer to the appetitive stimulus, irrespective of what appetitive stimulus is being approached. As such, there is no need for this approach behaviour to be informed by anything other than the positive value of the to-be-approached stimulus, and thus a combination of a simple proximity reduction mechanism with $v^{\text{MF}}(s)$ is sufficient to account for approach. Similar arguments can be made for at least some species-specific aversive responses (Seligman, 1970).

However, the bare model-free value $v^{\text{MF}}(s)$ alone cannot account for what to do with the appetitive stimulus, i.e. for consummatory behaviours. A positive $v^{\text{MF}}(s)$ indicates that reward is expected, but not whether it will require chewing (for a pellet), licking (for water), or copulation (for a sexually receptive conspecific). In order to produce such consummatory behaviour the model-free value must modulate, or somehow be informed by, a system that has access to the relationship between responses and outcomes or stimuli (Rescorla and Solomon, 1967). Note that such learned consummatory responses can be elicited in parallel with the simpler approach behaviour. As action-outcome representations are central to the notion of model-based systems, it is likely that consummatory responses, and indeed the transfer of consummatory responses to stimuli (Davey and Cleland, 1982) arises from a modulation of a (possibly evolutionarily restricted) model-based system by model-free values akin to Pavlovian-Instrumental transfer (see below). There is in fact evidence for a neural dissociation between approach and consummatory Pavlovian responses, with a certain alignment with model-based and model-free circuits (Yin et al., 2008), although the interaction between these is not clear. However, not all putatively consummatory responses adaptively reflect actions that are adapted to the US (Hearst and Jenkins, 1974).

4.3 INSTRUMENTAL BEHAVIOUR

Despite not containing information about actions, model-free Pavlovian values $\mathcal{V}^{\text{MF}}(s)$ can drive the acquisition of instrumental behaviours via multiple paths. The acquisition of $\mathcal{V}^{\text{MF}}(s)$ is based on bootstrapping, iteratively updating estimates of the value to fit with the sum of the current reward and the value of the next state. In this process, the cached value $\mathcal{V}^{\text{MF}}(s)$ comes to *replace* the summed future rewards. More specifically, changes in state values $\mathcal{V}^{\text{MF}}(s)$ imply changes in future reward, and so a change in value induced by an action is a metric that can be used to reinforce behaviours. This forms the core of the actor-critic model (Barto et al., 1983; O’Doherty et al., 2004). Experimentally, it is perhaps most directly demonstrated by conditioned reinforcement experiments (Everitt and Robbins, 2005; Meyer et al., 2012), where instrumental behaviours can be reinforced by Pavlovian CSs.

Model-free values also can have other influences on model-based instrumental behaviour. Determination of model-based values $\mathcal{Q}^{\text{MB}}(s, a)$ often require too much computational power to be feasible, as we emphasized above. One powerful approach is to mix model-based and model-free evaluations, and this has been successfully used in building computers that beat world chess masters (Campbell et al., 2002). Returning to Figure 1A, such an approach would correspond to replacing the subtree below a particular node with that node’s model-free value. This thus forms a second path by which model-free Pavlovian state values can drive instrumental behaviour, and indeed by which model-free can drive model-based choices. Although such a subtree substitution is yet to be demonstrated experimentally, it is likely that drug seeking involves such a process: here, highly complex, circumspect and flexible behaviours facilitate approach to a drug (a tree search up to a particular node in the tree); but the negative consequences of taking the drug are not respected (the tree below the node is not evaluated).

4.4 PAVLOVIAN-INSTRUMENTAL TRANSFER

Both model-free $\mathcal{V}^{\text{MF}}(s)$ and model-based $\mathcal{V}^{\text{MB}}(s)$ Pavlovian values can influence instrumental behaviour. This is demonstrated in two types of Pavlovian-instrumental transfer (PIT), general and outcome-specific PIT. In both types of PIT, appetitive CSs enhance and aversive CSs suppress instrumental behaviours for other outcomes (Estes and Skinner, 1941; Rescorla and Solomon, 1967; Lovibond, 1983; Cardinal et al., 2002; Niv et al., 2007; Holmes et al., 2010; Talmi et al., 2008; Huys et al., 2011). In general PIT, a stimulus that has been paired in a Pavlovian manner with one type of outcome (e.g. water) increases instrumental lever pressing for another type of outcome, (e.g. pellets). The specific nature of the expected reward is not relevant, only its value. Hence, the information that is present in $\mathcal{V}^{\text{MF}}(s)$ values may be sufficient for this.

In contrast, in outcome-specific PIT, a CS associated with pellets promotes an instrumental action reinforced by pellets over and above another instrumental action that was reinforced by sucrose (Corbit and Balleine, 2005). This does require representation of the actual outcome, not just the value. Thus, while general PIT requires only the information carried by $\mathcal{V}^{\text{MF}}(s)$, outcome-specific PIT requires additional information and likely relies on $\mathcal{V}^{\text{MB}}(s)$ from model-based processes (Corbit and Balleine, 2005; Schoenbaum et al., 2009; Holmes et al., 2010; McDannald et al., 2011; Prévost et al., 2012, 2013).

4.5 MOTIVATIONAL SHIFTS

The temporal integration of prediction errors has one further important consequence: it instantiates a slow, running average over experience. This means that model-free systems will not immediately reflect changes in motivation. Model-based systems on the other hand will. Motivational shifts have been used to highlight model-based components in both Pavlovian and instrumental scenarios. We recall that prediction error signals have been found not only in Pavlovian, but also in instrumental scenarios (Morris et al., 2006; Roesch et al., 2007).

First, consider instrumental devaluation experiments. An animal is first trained to perform a re-

sponse, say press a lever, for a reward. The reward is then devalued, for instance by giving the animal free access to it followed by administration of a nausea-inducing drug. When given another opportunity to consume it, the animal will refuse to do so. If the animal has had extensive experience with the behaviour, then it will initially continue to press the lever despite refusing to consume the food. This habitual behaviour is said to be under stimulus-response control, and not under the control of a representation of the outcome (Dickinson and Balleine, 1994, 2002). In other words, the information reflected in the habitual behaviour is present in a stimulus-action value $Q^{MF}(s, a)$, which captures how valuable an action is in a particular state, but without providing any information about the actual outcomes. An insensitivity to motivational changes is characteristic of cached values and habitual choices (McClure et al., 2003b; Daw et al., 2005; Valentin et al., 2007; Daw et al., 2011; Wunderlich et al., 2012a). Thus, instrumental learning derived from the accumulation of dopaminergic prediction errors accounts for outcome-insensitive habits.

After less extensive instrumental training, animals *are* sensitive to devaluation, and a reduction of behaviour can be observed on the very first trial after devaluation (Dickinson and Balleine, 1994, 2002), suggesting that a prospective representation of the outcome of the action is used to guide action choice. Similar findings hold in closely related paradigms in humans (Valentin et al., 2007; de Wit et al., 2009; Tricomi et al., 2009; Daw et al., 2011). The shift from early devaluation sensitivity to later devaluation insensitivity can be explained by the statistical properties of model-based and model-free systems, respectively. The model-free system has comparatively poor accuracy when little data is available, but this improves with experience (Daw et al., 2005; Keramati et al., 2011). Motivational shifts appear to have less effects on actions proximal to the goal (Killcross and Coutureau, 2003; Daw et al., 2005) where the burden on tree search is low. One complication to this account is the requirement for incentive learning in certain situations. Animals trained hungry may not change their behaviour when tested thirsty unless they have experienced the outcomes in those particular motivational states. This suggests a certain inaccessibility of internal states to the model-based system, at least in instrumental settings.

Motivational shifts can also be used to demonstrate model-based components in Pavlovian conditioning (McDannald et al., 2011; Dayan and Berridge, 2013). A striking example was recently provided by Robinson and Berridge (2013), where animals were first trained to associate a CS with aversive Dead Sea salty water, such that presentations of the CS readily induced aversive responses. Strikingly, after rendering the animals hungry for salt, they immediately started approaching the CS. Thus, a motivational shift succeeded in rendering a previously aversive stimulus appetitive. Clearly, the rapid approach after the motivational shift cannot be accounted for by a cached stimulus value - this would require multiple iterations of sampling the salt water in the salt-hungry state before the new positive prediction errors could update the stimulus value sufficiently to make it attractive. Instead, this experiment suggests that the animals learned the identity of the outcome associated with the stimulus, and in the novel salt-hungry state were able to use this to *infer* the new value of the stimulus given the new value of the outcome it predicted (Schoenbaum et al., 2009; Jones et al., 2012; Dayan and Berridge, 2013).

4.6 UNLEARNING

Finally, in the model-free equation 7 the entire past reward history is contained in $V^{MF}(s)$. No other aspect of the past history is maintained, and the past values are forgotten as soon as a change occurs. Say, for instance, a CS has $V^{MF}(s) = 4$, predicting 4 pellets, but then is updated (via a prediction error) to $V^{MF}(s) = 5$. The latter is the only representation maintained; there is no memory of the fact that the CS used to predict less pellets in the past. Hence, any **learning** to reflect new information equivalently implies forgetting or **unlearning** past information (Bouton, 2004; Rescorla and Wagner, 1972). Although slow changes can indeed lead to unlearning, sudden shifts in the predictive validity of a stimulus (extinction learning) do not lead to unlearning but rather to the learning of novel associations (Bouton, 2004). Such novel associations correspond to the learning of new latent causes for observations (Courville et al., 2004; Courville et al., 2005; Gershman et al., 2010). Unlike unlearning, these fit more easily in a model-based than a model-free framework.

5 INDIVIDUAL VARIABILITY

We have now reviewed model-based and model-free learning, the role of dopamine in model-free learning, and behavioural and neurobiological characteristics of both systems. Recent findings have highlighted substantial individual variability in *how* and *what* subjects learn in standard Pavlovian conditioning paradigms. This has consequences for learning accounts of addiction as some learning tendencies appear to confer vulnerability towards developing addiction. In this part, we first present the data on individual differences in Pavlovian responding in some detail (mainly reiterating the findings of [Flagel et al. 2011b](#)), then discuss its interpretation in terms of incentive salience ([Berridge and Robinson, 1998](#); [Berridge, 2004, 2007](#); [Saunders and Robinson, 2012](#)), and finally put forth a hypothesis that proposes a connection between the propensity to assign incentive salience and the propensity to employ model-free learning ([McClure et al., 2003a](#); [Huys et al., 2013b](#); [Lesaint et al., 2013](#); [Dayan and Berridge, 2013](#)).

5.1 SIGN-TRACKING AND GOAL-TRACKING

5.1.1 Behaviour

When rats are exposed to a conditioned stimulus (CS), such as a lever, that is repeatedly paired in a Pavlovian fashion with an unconditioned stimulus (US), such as food reward, there is substantial individual variability in the conditioned response that emerges (see [Figure 3](#)). Some animals, referred to as “sign-trackers” (STs) will approach and oftentimes interact with the CS upon its presentation ([Figure 3A](#); [Hearst and Jenkins 1974](#)). Others, termed “goal-trackers” (GTs) approach the location of reward delivery upon CS presentation ([Figure 3D](#); [Boakes 1977](#)). Remarkably, these conditioned responses develop even though reward delivery is not contingent on any response, i.e. in a classical Pavlovian conditioning paradigm. Furthermore, all rats learn the CS-US association, the resulting conditioned responses emerge at similar speeds, and both STs and GTs retrieve all of the food pellets that are delivered. Hence, the topography of the emitted response differs, but both sets of animals learn the CS-US association equally well and at similar speed.

5.1.2 Dopamine signals during acquisition

These individual differences in conditioned responding have shed light on the role of dopamine in stimulus-reward learning. [Flagel et al. \(2011b\)](#) used fast-scan cyclic voltammetry in the core of the nucleus accumbens to characterize cue-induced phasic dopamine signaling during Pavlovian training in selectively bred rats predisposed towards sign- or goal-tracking behavior. Similar to outbred rats, these selectively bred phenotypes both learned a conditioned response and did so at the same rate. Further, the lever-CS was more attractive and more desirable for the selectively bred sign-trackers, as indicated by approach behavior and the ability of the lever-CS to serve as a conditioned reinforcer. Remarkably, only for sign-trackers did the lever-CS evoke an increase in dopamine release. That is, only for sign-trackers did the phasic dopamine response shift from the presentation of the food reward-US to the lever-CS across training. The CS did evoke dopamine release in goal-trackers, but this did not change over trials. The same pattern of results was also found in outbred rats characterized as sign- or goal-trackers, suggesting that these neurochemical signatures are specific to the conditioned responses and not an artifact of the selective breeding.

Next, [Flagel et al. \(2011b\)](#) asked whether the development of either a goal- or sign-tracking responses was dependent on dopamine. They administered a nonspecific dopamine antagonist, flupenthixol systemically prior to the first of several Pavlovian training sessions. The selectively-bred animals were ideal for this experiment, as the predictability of the phenotypes allowed the authors to assess the effects of this drug on the acquisition of the conditioned responses. Interestingly, administration of the dopamine antagonist attenuated the performance of both sign-tracking and goal-tracking behavior. However, when taken off of the drug, the goal-trackers exhibited a fully developed conditioned response, similar to control animals; whereas the sign-trackers remained deficient in their responding even during the drug-free test session. Thus, dopamine was necessary for learning the CS-US association in sign-trackers, but not in goal-trackers. Similarly, [Parker et al. \(2010\)](#) reported that mice with disrupted dopamine signaling were fully capable of learning

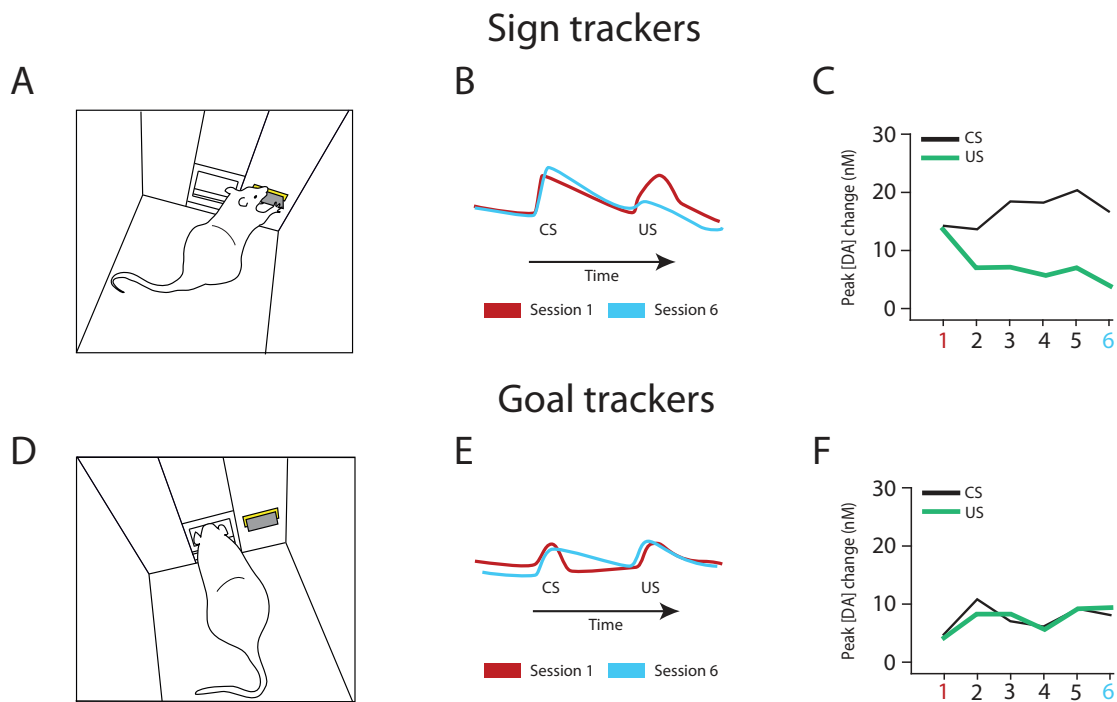


FIGURE 3: Sign tracking and goal tracking by animals exposed to classical conditioning, whereby a CS (in the figure a lever on the right) predicts delivery of a food US at a different location (in the food box on the left). Note that this is a Pavlovian conditioning procedure, and thus the rat obtains the food irrespective of its behaviour, and does not need to press the lever. **A&D**: Sign tracking rats come to approach the lever CS during CS presentation, while goal-tracking rats approach the location where the food US will be delivered. **B&E**: Phasic dopamine signals in the nucleus accumbens core. In sign trackers, the phasic response to the CS increases, while that to the US decreases, as is predicted by the temporal prediction error hypothesis. In goal-trackers, phasic dopamine responses to CS and US do not change over time. **C&F** show how the peak dopamine responses change over trials. These differences suggest that sign trackers acquire a cached value $\mathcal{V}^{MF}(s)$ in accordance with the temporal prediction hypothesis, but that goal trackers do not. Data in B,C,E,F adapted from Flagel et al. (2011b).

a goal-tracking conditioned response, despite the fact that there was no transfer in dopamine signaling from the US to the CS. Thus, phasic dopamine signals are critical for learning the CS-US relationship that leads to a sign-tracking conditioned response; but not for those that lead to a goal-tracking conditioned response.

5.1.3 Dopamine signals after acquisition

Flagel et al. (2011b) also examined the effects of flupenthixol on the expression of sign- and goal-tracking behavior after the conditioned responses were acquired. They found that systemic dopamine antagonism attenuated the expression of both. To more directly assess the role of dopamine in the performance of these conditioned behaviors, and to minimize non-specific effects of the drug on behavior, Saunders and Robinson (2012) administered flupenthixol directly into the core of the nucleus accumbens after outbred rats had acquired stable sign- or goal-tracking behavior. This dose-dependently attenuated sign-tracking behavior, with little to no effect on goal-tracking behavior (see also Di Ciano et al. 2001; Parkinson et al. 2002). Importantly, sign-tracking behavior was fully impaired upon the first CS-US presentation following administration of flupenthixol into the accumbens. Thus, the drug effects were evident before new learning could occur, and changes in dopamine levels were able to alter the motivational value of reward cues,

without the need to re-experience the CS-US association (Berridge, 2012; Richard et al., 2013; Dayan and Berridge, 2013; Robinson and Berridge, 2013). Furthermore, the effects of dopamine antagonism were specific to the Pavlovian conditioned approach behavior and did not affect the conditioned orienting response in the sign-trackers (Saunders and Robinson, 2012).

5.2 INCENTIVE SALIENCE ACCOUNTS OF THE SIGN-TRACKING/GOAL-TRACKING VARIABILITY

Attribution of incentive salience is the process by which neutral stimuli are transformed into attractive and “wanted” incentive stimuli via Pavlovian learning mechanisms (Berridge, 1996; Berridge and Robinson, 2003). Extensive research has shown that Pavlovian stimuli that have been attributed incentive salience have three fundamental properties (Berridge, 2012): 1) they are attractive and elicit approach towards them, 2) they are themselves desirable and can reinforce the learning of new actions (i.e. act as conditional reinforcers), and 3) they can elicit a conditioned motivational state, energizing ongoing instrumental actions (i.e. general Pavlovian-to-instrumental transfer (PIT); Cardinal et al. (2002); Everitt et al. (2001); Milton and Everitt (2010)). These three features are dissociable, but rely on partially overlapping neural mechanisms (Cardinal et al., 2002). Incentive salience in this context is distinct from incentive motivational properties or “incentive value” in instrumental settings as defined by Dickinson and colleagues (Dickinson and Balleine, 1994; Dickinson et al., 2000).

5.2.1 Behaviour

The incentive salience account of sign-tracking/goal-tracking describes the difference between the two groups, arguing that CSs are imbued with incentive salience by sign-trackers, but not by goal-trackers. Both sign-trackers and goal-trackers learn that the lever-CS precedes and predicts the delivery of the US in that the lever-CS comes to elicit a response in both phenotypes, and respective responses emerge at a comparable rate. As they emit their response similarly, and this response has the same relationship to the predicted US (i.e. it is non-contingent), both phenotypes are equally able to assign “predictive” value to the CS. However, only for sign-trackers does the lever-CS attain the additional incentive motivational properties mentioned above. Hence, the assignment of incentive value is seen as the central component that distinguishes sign-trackers and goal-trackers. The ability of the CS to predict the occurrence of the US is considered to be common to both groups.

For sign-trackers, the CS attains at least two of these fundamental properties of an incentive stimulus (i.e. of a stimulus that has acquired incentive salience), which is reflected in the nature of the conditioned responses (Robinson and Flagel, 2009). First, sign-trackers (unlike goal-trackers) approach the conditioned stimulus upon its presentation, and the cue is attractive to them. Second, sign-trackers exert more instrumental effort than goal-trackers for presentation of the CS in the absence of food reward. Thus, the cue acts as a more powerful conditioned reinforcer for sign-trackers than for goal-trackers (Robinson and Flagel, 2009; Lomanowska et al., 2011; Meyer et al., 2012). Evidence demonstrating individual variation in the third fundamental property of an incentive stimulus, i.e. general PIT, is lacking, perhaps due to the complex nature of the paradigm. Taken together, these findings support the notion that for sign-trackers, but not goal-trackers, the lever-CS is attributed with incentive salience. Salience attribution theories hence consider the assignment of incentive value to be the central component that distinguishes sign-trackers from goal-trackers.

5.2.2 Dopamine

In the incentive salience framework, dopamine is specifically involved in assigning Pavlovian incentive value in sign-trackers. This relies on a tight link between the three key features of incentive salience reviewed above, and dopamine. First, the shift in phasic responses from the US to the CS is present in sign-trackers but not in goal-trackers (Flagel et al., 2011b). As both groups learn the association between CS and US, but differ in terms of the gradual attribution of incentive salience, this suggests that phasic dopamine is relevant not to learning to predict the US from the CS *per*

se, but to assigning incentive value to the CS. Second, non-selective dopamine antagonism affects learning of a sign-tracking response, but it does not affect learning of a goal-tracking conditioned response (Flagel et al., 2011b). This complements the findings demonstrating a selective shift in phasic responding in sign-trackers, but not goal trackers, and argues that: 1) dopamine is necessary for the assignment of incentive salience and 2) dopamine is *not* involved in the assignment of the predictive properties that are also seen in goal-trackers. Third are the results from injections of dopamine antagonists into the nucleus accumbens core after completion of learning. These have immediate effects, before any new learning can occur. This suggests a role for dopamine in incentive salience that goes beyond that of learning. Furthermore, the fact that orienting responses were unaffected (in both sign- and goal-trackers) suggested that even in sign-trackers, dopamine was not abolishing all the of the qualities of the CS; only its incentive salience properties (i.e. its ability to elicit approach). Finally, dopamine abolished the conditioned response in sign-trackers only (Saunders and Robinson, 2011), which again argues for a role that is selectively associated with incentive salience processes. Hence, it appears clear that dopamine has an involvement in incentive salience that is independent of and goes beyond its involvement in learning; and that some aspects of learning the CS-US associations remain intact in the absence of dopamine, not only in goal-trackers, but in sign-trackers, too.

5.3 REINFORCEMENT LEARNING ACCOUNTS OF THE SIGN-TRACKING/GOAL-TRACKING VARIABILITY

We now consider the hypothesis that model-free and model-based *learning* may at least partially map onto sign- and goal-tracking behaviour, respectively (Lesaint et al., 2013; Huys et al., 2013b). In Pavlovian conditioning experiments, reward delivery is independent of the animals' behaviour. Hence, only stimulus, but not stimulus-action values, are constrained. RL accounts match incentive salience accounts in terms of arguing that dopamine is relevant for sign-trackers but not goal-trackers, but differ in a number of important details.

5.3.1 Behaviour

Section 4 detailed the characteristics of behaviour that model-free values *can* and *cannot* support. The suggestion that incentive salience, and hence sign-tracking, is driven by $v^{MF}(s)$ values hinges on arguing that model-free values $v^{MF}(s)$ are sufficient to account for the three fundamental properties of incentive salience (see section 4); and that the behaviour shown by sign-trackers does not require access to information that cannot be contained in $v^{MF}(s)$. This is because $v^{MF}(s)$ values are devoid of anything but the size of the expected reward. On their own, they can only influence behaviour as a 'pure' reward would *because* they represent no other information. Specifically, we make reference to the three key components of incentive salience mentioned in section 5.2. First, model-free values can drive Pavlovian approach responses (section 4.2). As such, they capture the key feature of sign-trackers as compared to goal-trackers. Second, they can reinforce actions as conditioned reinforcers are formalized in the actor-critic models (section 4.3). This captures the notion that stimuli assigned incentive salience can become conditioned reinforcers. Third, they can influence ongoing behaviour arising in other systems by altering the opportunity costs (section 4.4). This captures the ability of stimuli with incentive salience to influence other behaviour in general PIT experiments. However, these different features of incentive salience are known to have only partially overlapping neurobiological substrates. Similarly, for model-free values to lead to these features, they would have to interact with other systems (e.g. with instrumental systems both for conditioned reinforcement and PIT), and hence again only have partially overlapping neurobiological substrates. Nevertheless, $v^{MF}(s)$ values appear sufficient to account for the main features of sign-tracking behaviour and incentive salience (see also McClure et al. 2003b; Dayan and Berridge 2013).

5.3.2 Dopamine signals during acquisition

The parallel between model-free systems and sign-tracking is strengthened by the role of dopamine. Both sign-trackers and goal-trackers show phasic DA responses in the NAcc core to both CS and US onsets (Figure 3B,E, red traces). In the sign-trackers this signal changes slowly over time, in-

creasing in response to the CS and decreasing in response to the US (Figure 3B, blue trace and C). As extensively reviewed in section 3, this is what would be expected *if* the prediction error was based on the slow, iterative, accumulation of a cached value $v^{\text{MF}}(s)$. In sign-trackers, interfering with these dopamine signals by injecting a non-selective dopamine antagonist during training prevents any learning (Flagel et al., 2011b; Parker et al., 2010), which is in keeping with results on other Pavlovian behaviours such as autoshaping (Di Ciano et al., 2001; Parkinson et al., 2002) and mirrors the findings that phasic dopamine signals can have a causal role in Pavlovian learning (Steinberg et al., 2013). It suggests, thus, that sign-trackers *need* a phasic dopaminergic prediction-error signal in order to learn *because* their learning is heavily biased towards learning through incremental acquisition of model-free values $v^{\text{MF}}(s)$. The fact that the signals are observed in the NAcc core also maps onto the notion that these signals might be model-free because, as discussed above, model-free mechanisms suffice for general PIT, which is dependent on the core, but not for specific PIT, which is more dependent on the shell (section 4.4; Corbit and Balleine 2011; though see (Shiflett and Balleine, 2010; Robinson and Berridge, 2013)). Finally, the reliance on model-free learning can, at least in part, explain the core incentive salience features.

5.3.3 Dopamine signals after acquisition

The results by Saunders and Robinson (2012) clearly suggest that the role of dopamine is not limited to representing phasic error signals for learning, but extends to the expression of behaviour once learning has stabilised (see also Shiner et al. 2012). One likely and important issue is that Saunders and Robinson (2012) manipulated not only phasic but also tonic dopamine signals. Indeed, the most prominent effects of manipulations of dopamine are not alterations in learning, but profound changes in the rate and vigour at which behaviour is emitted (Salamone et al., 2009). The reinforcement learning framework reviewed above does not account for this, but semi-Markov, average reinforcement learning formulations do (Niv et al., 2007). These consider not only *which* action to emit, but also *when* and *how vigorously*. They achieve this via an extra term, the average reinforcement, which functions as an opportunity cost (i.e. as a measure of reward forfeited on average by inaction). Examinations of the impact of this term on behaviour suggested a close link with tonic dopamine (Niv et al., 2007). This could potentially explain the impact of dopamine antagonists on the *expression* of both sign- and goal-tracking behaviour during learning (Flagel et al. 2011b; see also Mazzoni et al. 2007; Beierholm et al. 2013).

The results of Saunders and Robinson (2012) however show that after learning the impact of dopamine antagonists is confined to sign-trackers. Interpreted in the RL framework, this suggests that the opportunity cost might be preferentially mediated via tonic dopamine in those animals that rely on model-free learning whereas the timing and vigour of model-based choices might be more directly linked to the anticipated outcome, and hence less sensitive to such tonic dopaminergic mechanisms. Indeed, interference with DA by pharmacological means or by VTA inactivation abolish the ability of Pavlovian CSs to motivate approach and produce PIT (Wassum et al., 2011; Murschall and Hauber, 2006; Lex and Hauber, 2008), and DA stimulation promotes it (Wyvell and Berridge, 2000); whereas model-based behaviour is often rather more resilient to DA manipulations (e.g. Wassum et al. 2011; though see Wunderlich et al. 2012b; Guitart-Masip et al. 2013). Thus, the admittedly very speculative suggestion is that tonic levels of dopamine in the NAcc core differentially modulate the expression of model-free values, and thereby differentially affect sign-trackers.

5.3.4 Goal-trackers

The RL account of goal-tracking behaviour is less crisp, both theoretically and in terms of its mapping onto neurobiological substrates. As pointed out earlier, goal-trackers clearly make predictions about the occurrence of rewards as they are perfectly able to approach the goal-box upon presentation of the CS. As explained in the previous section, predictions of reward associated with stimuli can be derived not only from model-free ($v^{\text{MF}}(s)$), but also from model-based ($v^{\text{MB}}(s)$) learning. Indeed, that is the very *raison d'être* for both, and so the fact that both sets of animals make predictions is not informative about which mechanism they learn by. More to the point, $v^{\text{MF}}(s)$ values are sufficient to produce both the “predictive” and “incentive” learning. However, the fact that the CS is itself less attractive and supports less conditioned reinforcement in goal-trackers

suggests that it has not acquired features of a reward itself (as model-free values do), but rather helps the rat explicitly predict that a particular event (a reward, in this case) will happen in the future. Model-based learning of \mathcal{T} might consist in learning to predict that the event 'CS' is followed by the event 'pellet delivery' (i.e. the statistical rules of the environment), while the structure \mathcal{R} would separately be used to represent the desirability of that event. There is evidence that signals involved in acquiring \mathcal{T} are differentiable from reward prediction error signals (Gläscher et al., 2010). Thus, when seeing the CS, a model-based learner in the autoshaping experiment might be reminded specifically of the food pellet, and base its action choice on its current desires; and the learning of this type of prediction appears not to depend on dopaminergic prediction errors. The CS would be a purely "informational" stimulus, not attractive in its own right (Flagel et al., 2011a). This account makes a very straightforward and easily tested prediction, namely that food devaluation should abolish goal-tracking, but leave sign-tracking unchanged. This is at least partially consistent with reports whereby highly deprived animals (at 75% of optimal body weight) show stronger goal-directed behaviour than animals that are less deprived (at 90% body weight; Boakes 1977). However, this is certainly also consistent with effects motivation could have via goal-directed mechanisms, and indeed may be complicated by issues related to incentive learning.

By arguing that goal-trackers are indeed more goal-directed, this suggests the involvement of goal-directed neural structures (Yin et al., 2004, 2005; O'Doherty et al., 2004; Killcross and Coutureau, 2003). In agreement, goal-trackers do seem to recruit cortical "top-down" regulation of their response to reward-cues (Flagel et al., 2011a). This, however, then raises to question about the nature of the phasic dopamine signals in the goal-trackers. There clearly are phasic DA responses to both CS and US in the goal-trackers, but these stay constant without showing any signs of adaptation (Figure 3F). Areas thought to be involved in model-based Pavlovian estimation of values are known to influence phasic dopamine signals (Takahashi et al., 2011). However, as the size of the signals does not change, it suggests that the prediction term used in their computation must remain at zero, and hence that the prediction errors are not iteratively collated into a model-free value. Why would this be? There are several potential answers. It might be that the model-free system learns only 'on-line', i.e. only when it is in charge itself (Sutton and Barto, 1998). That this might be neurobiologically plausible is suggested by the fact that habitual control of behaviour is itself under constant control of the prefrontal cortex, specifically the infralimbic cortex (Smith et al., 2012; Smith and Graybiel, 2013). It might also be that the dopamine transient signals the need to change one's beliefs (i.e. *that* learning is necessary); but is not a teaching signal itself (i.e. does not indicate *what* should be learned; see also section 3.2). However, it is unclear why this signal would then continue to persist in animals after behaviour has reached a stable asymptote. A somewhat different explanation focuses on the detailed temporal structure of events, which differs between goal-trackers and sign-trackers. Goal-trackers focus on the goal as soon as the sign appears, but they also focus on the goal during the ITI (a time when the sign is not present) when no food is presented there. This may lead to keeping the model-free values of both the goal and the CS near zero (though ITI head-entries into the food-cup do not differ between goal- and sign-trackers; see Lesaint et al. 2013 for a detailed discussion).

In summary, reinforcement learning accounts might suggest that the "predictive" learning seen in goal-trackers is not dopamine dependent and relies on building a model of the structure of the environment. Conversely, it would suggest that the assignment of "salience" is evidence for relying on model-free learning via dopaminergic mechanisms.

6 ADDICTION

Addiction is a disorder with profound deficits in decision-making. Most addictive drugs have rapid effects and impact the dopaminergic system either directly or indirectly (Koob, 1992; Olds, 1956; Tsai et al., 2009). Several features of addiction are at least partially amenable to explanations within the overall framework outlined above. We will briefly consider partial accounts of addiction based on a) drug-induced alterations to phasic dopaminergic signals and b) individual (and drug-induced) variation in the tendency to rely on model-free learning and assign incentive salience

(Redish, 2004; Redish et al., 2008; Dayan, 2009; Flagel et al., 2011b; Huys et al., 2013b).

6.1 PHASIC DOPAMINERGIC SIGNALS IN ADDICTION

If drugs of abuse alter or directly elicit phasic dopamine release (Boileau et al., 2003, 2007; Cox et al., 2009), they could elicit artificial prediction errors which in turn would lead to enhanced learning of stimuli that predict their occurrence (Redish, 2004; Dayan, 2009). Indeed, L-Dopa enhances striatal prediction errors and learning whereas haloperidol reduces them (Pessiglione et al. 2006 though see also Knutson and Gibbs 2007). If drugs of abuse mimic dopamine prediction error signals, resulting in an irreducible, constant prediction error even in the absence of reward, then this would lead to a never-ending increase of the associated state $V^{MF}(s)$ or state-action $Q^{MF}(s, a)$ values, which would lead to strongly determined behaviour that would be hard to overcome. Blocking paradigms (Kamin, 1969; Waelti et al., 2001; Steinberg et al., 2013) provide one formal test of this prediction: new stimuli added to pretrained stimuli should be learned more if the reward used is a drug of abuse than if it is a natural reward (Redish, 2004). Administration of d-amphetamine into the nucleus accumbens enhances blocking in an aversive paradigm, whereas administration of dopamine antagonists reduce blocking (Iordanova et al., 2006). The prediction has also been tested explicitly for nicotine (though the results are, to our knowledge, only present in abstract form; Jaffe et al. 2010) and for cocaine (Panlilio et al., 2007). While in the former case they have been at least partially confirmed, by way of individual variation (highly nicotine responsive animals show no blocking for nicotine whereas animals more responsive to water do show blocking for water), the latter case failed to confirm this prediction. As pointed out by Dayan (2009), alternative forms of reinforcement learning that rely on actor-critic learning may allow for correct values (and hence blocking) despite a constant increase to prediction errors, and an effect directly on the advantage of actions could lead to more rapid development of deeply embedded actions, again with correct values.

Addiction is characterised by a profound and long-lasting downregulation of dopamine D₂ receptors in the striatum (Heinz et al., 1996, 2009; Volkow et al., 2009; Huys et al., 2013b), which is also characteristic of animal models of obesity (Johnson and Kenny, 2010). This downregulation may be a consequence of drug-taking, but it may also predispose to the development of addiction and to relapse (Volkow et al., 2009; Morgan et al., 2002; Thanos et al., 2001; Heinz et al., 2005; Volkow et al., 2002; Buckholtz et al., 2010). Dopamine D₂ receptors are both pre- and postsynaptically located. It is not clear whether the reduction seen in addiction is mainly pre- or postsynaptic, but both could potentially promote drug-taking. Postsynaptically, they have been shown to mediate the effect of losses on “go/no-go” learning (Frank, 2005; Frank et al., 2004; Dreyer et al., 2010; Kravitz et al., 2012), and could thereby contribute to the insensitivity towards adverse consequences in addiction (Vanderschuren and Everitt, 2004; Deroche-Gamonet et al., 2004; Maia and Frank, 2011; Kravitz et al., 2012). Presynaptically, they are involved in an autoinhibitory negative feedback loop which could particularly affect go-learning as it could reduce the positive phasic transients (Bello et al., 2011) and thereby lead to the sort of increased prediction error increase mentioned above (Bello et al. 2011; see also Sulzer 2011). Furthermore, drug-craving is correlated with the reduction in D₂ receptors (Heinz et al., 2004). It is conceivable that reductions in presynaptic D₂ receptors might also affect tonic dopamine signals (Martinez et al., 2005, 2009) and that this relates to the effects of dopamine and cached values on Pavlovian-instrumental transfer (Murschall and Hauber, 2006; Wyvell and Berridge, 2000) and sign-tracking (Saunders and Robinson, 2012). There is also evidence that the link between dopamine synthesis and phasic prediction errors is altered by addiction, and this might be mediated by a failure of the presynaptic D₂ control (Schlagenhauf et al., 2013; Deserno et al., 2013). Moreover, Flagel et al. (2010, 2013) have shown that selectively bred rats with a predisposition towards sign-tracking behaviour and addiction have lower levels of D₂ mRNA in the nucleus accumbens and dorsal striatum, but not in the ventral tegmental area. However, these addiction-prone rats also exhibit a greater proportion of striatal “D₂-high” receptors, the functionally active state of the dopamine D₂ receptor.

Thus, there is substantial theoretical and biological plausibility supporting the notion that drugs of abuse directly interfere with phasic dopaminergic signals, and that this contributes to the establishment and possibly to the maintenance of addicted behaviour. It has to be noted that it is

unclear, as yet, whether changes in dopamine signalling are a cause or consequence of drug abuse; although some of the animal literature suggests it may be a predisposing factor (e.g. Flagel et al. 2010; Dalley and Everitt 2009). Direct tests of this hypothesis have at present provided only equivocal evidence but these findings may be in part confounded by variability in the innate tendency of individuals to rely on model-free learning and assign incentive salience.

6.2 INDIVIDUAL VARIABILITY IN ADDICTION VULNERABILITY

As discussed above, there is growing evidence that the natural tendency to sign-track is both highly variable (Meyer et al., 2012) and a risk factor predisposing to addiction (Saunders and Robinson, 2010; Saunders et al., 2013a,b). That is, individual variation in cue reactivity is associated with individual differences in vulnerability to addiction. Rats that sign-track to cues associated with food reward also sign-track to drug-associated cues (Flagel et al., 2010); and drug cues maintain drug self-administration and reinstate drug-seeking behaviour to a greater extent in sign-trackers than goal-trackers, even in the face of adverse consequences (Saunders and Robinson, 2011, 2012; Saunders et al., 2013a). Furthermore, sign-trackers also express other traits related to addiction liability. They are, for instance, more impulsive than goal-trackers (Flagel et al., 2010; Lovic et al., 2011; Tomie et al., 1998) and more likely to seek novel environments (Beckmann et al., 2011). Finally, differences in the dopamine system have been associated with individual variation on all of these traits (Dalley and Roiser, 2012; Flagel et al., 2009). As we have argued that sign-tracking reflects incentive salience and model-free processes, this further motivates the suggestion that variations in the extent to which individuals rely on model-free learning processes form a risk factor for addiction.

Likewise, there is considerable individual variation in the ability of drug cues to bias attention, elicit craving and instigate relapse in humans (de Wit et al., 1986; Carter and Tiffany, 1999). Emerging evidence suggest that some humans may be more “cue reactive” than others. For example, Mahler and de Wit (2010) reported that individuals with the highest craving in response to food cues, when hungry, were the same individuals that showed the highest craving in response to smoking cues during abstinence. These findings are reminiscent of the sign-tracking rats that attribute excessive incentive motivational value to both food- and drug-paired cues (Flagel et al., 2011b; Saunders and Robinson, 2011, 2012) and lend credence to the notion that variation in this trait may underlie susceptibility to addiction in humans.

Just as in animals, this variation may be related to the dopaminergic system (Buckholtz et al., 2010; Dalley and Roiser, 2012). Leyton et al. (2002) showed that even in healthy subjects the variability in dopamine response to amphetamine relates to subjective ratings of “wanting”. Franken et al. (2004) showed that the dopamine receptor antagonist haloperidol can reduce attentional bias to drug cues amongst addicts, and Ersche et al. (2010) showed that the effect of such dopaminergic manipulations (both agonistic and antagonistic) varies with compulsivity. The effect of haloperidol in the former and amisulpiride in the latter, both rather selective D₂ antagonists, is surprising given the drug-induced reductions in D₂ receptors (see above). However, the directionality of the effect is consistent with the pro-compulsive and pro-addictive effects of D₂ agonists in Parkinson’s disease and may relate to specific effects in the ventral compared to the dorsal striatum (Evans et al., 2006; Dagher and Robbins, 2009).

Although we have focused on evidence from Pavlovian learning (particularly sign-tracking), the reliance on and shift towards model-free learning is also apparent in instrumental learning, with addictive drugs shifting responding from model-based towards model-free responding, speeding up habitization and likely predisposing towards addiction. When rats acquire instrumental responses for alcohol they become insensitive to devaluation earlier than when the outcome is pellets (Dickinson et al., 2002). Along the same lines, amphetamine pre-treatment speeds up the rate at which the outcome insensitivity develops (Nelson and Killcross, 2006), and this depends particularly on D₁ rather than D₂ receptors (Nelson and Killcross, 2013). In humans, there is evidence for enhanced habitization in obsessive-compulsive disorder (Everitt and Robbins, 2005; Robbins et al., 2012; Gillan et al., 2011, 2013) and forthcoming evidence in cocaine addiction (N. Daw and V. Voon, personal communication), but not yet in alcohol addiction (Sebold et al., *subm.*).

Both innate variability in attributing incentive salience and relying on model-free learning, and more direct effects on dopaminergic signals predict that drug-associated cues should have increased model-free value in addicts. This in turn means that sudden, unexpected presentation of such cues should elicit greater dopaminergic transients. PET studies measuring released dopamine with raclopride displacement (Kalivas and Volkow, 2005; Volkow et al., 2006; Boileau et al., 2007) and fMRI studies measuring responses to drug-associated cues (Grüsser et al., 2004; Wrase et al., 2007; Beck et al., 2009) both clearly support this prediction (though see Wilson et al. 2004 for a discussion of how these relate to craving).

Finally, it is worth emphasizing the impact of past experience on present learning (Huys et al., *subm*). A stimulus that elicits approach will be more attended to, and hence may be more easily learned about and associated with reinforcements at the expense of other stimuli present in the environment. More generally, on-line iterative reinforcement learning in which behaviour (and hence sampling of the environment) changes after every experience often does not have the kind of optimality guarantees that off-line learning has (Bertsekas and Tsitsiklis, 1996), and may lead to self-reinforcing loops of choice and reward (Hogarth et al., 2007). One such effect was shown directly by Freeman et al. (2012), who have found that abstinent smokers were more likely to associate a drug cue with reward than a non-drug cue. Indeed, attentional mechanisms are clearly important in learning (Pearce, 1997; Dayan et al., 2000) and possibly in the maintenance of addiction (Hogarth and Chase, 2011; Hogarth et al., 2013; Wiers et al., 2011).

6.3 SHIFTS TOWARDS MODEL-FREE LEARNING IN ADDICTION

We have so far mainly focused on contributions by the model-free system. However, alterations to the model-based systems are likely to be equally important, and open alternative paths to addiction. As reviewed above, extinction does not lead to unlearning, but rather to the re-engagement of prefrontal cortices and novel learning (Bouton, 2004; Gershman et al., 2010). The underlying associations continue to be present and can re-emerge, either spontaneously or in response to a cue. Interestingly, context-induced reinstatement is more prominent in goal- than in sign-trackers (Saunders and Robinson, 2013). Moreover, a context paired with ethanol injections can immediately and profoundly impair the ability to exert goal-directed control (Ostlund et al., 2010), and optogenetic suppression or activation of the prelimbic cortex, which is thought to involve goal-directed computations, can abolish or re-establish sensitivity to punishments (Chen et al., 2013). There is also preliminary evidence for this in humans (Sebold et al., *subm*). Addiction, therefore, might impair the normal re-engagement of model-based decision-making in the face of aversive events or events triggering drug-related behaviours (Ostlund and Balleine, 2008). The former may account for the perseverance of behavioural response patterns in the face of adverse consequences, a hallmark of addiction (Gelder et al., 2006; Vanderschuren and Everitt, 2004; Deroche-Gamonet et al., 2004).

In a landmark study, Killcross and Coutureau (2003) showed that lesions of the pre- and infralimbic rodent cortices abolished goal-directed and habitual behaviour, respectively. This showed that model-free and model-based systems co-exist in the brain, but that behavioural expression tends to be dominated by one or the other. Behavioural and imaging evidence for this also exists in humans (Daw et al., 2011). This immediately raises the question of arbitration: how is dominance determined? There are two prominent explanations. Daw et al. (2005) argued from a Bayesian perspective that it would be optimal to use all knowledge when making choices, but that various types of knowledge should be weighted by their certainty. Using detailed analyses of the noise characteristics of model-based and model-free systems, they argued that model-based systems are more data-efficient, and hence make more accurate predictions, when little evidence exists and uncertainty is high, i.e. early on in training. The opposite is true after extensive evidence later in training. An alternative account (Keramati et al., 2011; Pezzulo et al., 2013) is based on the value of information (VOI; Russell and Wefald 1991). Unlike the Bayesian account, this explicitly takes the cost of computation into account. Briefly, if the expected improvement in performance outweighs the cost of computation, then it is worth engaging in model-based reasoning. Because of a similar argument about the increasing accuracy of model-free values with experience as used by (Daw et al., 2005), this improvement is worthwhile early on in training, but not later on. However,

the VOI account is fundamentally different, in that it suggests that expression of habits is under continuous evaluation and control by the prefrontal cortex, which is consistent with some recent evidence (Cavanagh et al., 2013; Smith et al., 2012; Smith and Graybiel, 2013).

Both of these models provide multiple avenues for a shift from goal-directed to habitual behaviour. Both the VOI and the uncertainty-based account would increase the prominence of the model-free systems as a consequence of increased noise in the model-based system. In the former case, the increase in information that would occur from engaging the model-based system would be reduced. This could occur due to a general cognitive impairment (D. Schad, M. Rapp and Q. Huys, unpublished observations), perhaps due to deficits in prefrontal function, especially as a result of exposure to neurotoxic substances such as alcohol or cocaine (Goldstein et al., 2004; Briand et al., 2008; Lucantonio et al., 2012) but may also be characteristic of other populations (e.g. Darke et al. 2000), and does involve the prefrontal cortex (Goldstein et al. 2004; see also Volkow et al. 2009). In support, Takahashi et al. (2011) have recently shown that cocaine interferes with the ability of the orbitofrontal cortex to establish detailed state-space (Walton et al., 2010), which would lead to less accurate models and hence less accurate predictions. In the VOI account, increased cost of computation would have very similar effects, and cognitive impairments could be involved in the effect of stress (Schwabe and Wolf, 2009) and certainly the effect of dual tasks (Otto et al., 2013).

It is important, though, to bear in mind that although habits share features with compulsions (Gillan et al., 2011, 2013), they are not one and the same (Dayan 2009; Robbins et al. 2012 and many others). It has been suggested that after extended training, habits become deeply engrained by shifting further dorsally in the corticostriatal loops (Belin and Everitt, 2008; Willuhn et al., 2012). Using cyclic voltammetry and a behavioural paradigm similar to sign-tracking paradigm, Clark et al. (2013) examined changes in dopamine release in the nucleus accumbens core during the acquisition and maintenance of a Pavlovian conditioned approach response (i.e. sign-tracking). In agreement with the results of Flagel et al. (2011b), it was shown that contact with the lever-CS and CS-evoked dopamine release increased over time for rats that sign-tracked. However, after prolonged training (i.e. around 150 CS-US trials), these two measures were no longer correlated. That is, sign-tracking behaviour continued at asymptotic levels, but CS-evoked dopamine release diminished with extended training. Moreover, the effects of a dopamine D1 receptor antagonist on sign-tracking behaviour were less prominent following post-asymptotic training. However, the data on punishment sensitivity (Vanderschuren and Everitt, 2004; Deroche-Gamonet et al., 2004) and the importance of prefrontal mechanisms in the reassertion of control (Chen et al., 2013; Ostlund and Balleine, 2008; Sebold et al., *subm*) may also speak to the difference between habits and compulsions.

6.4 CONCLUSIONS

In this paper, we have suggested that the combination of a theoretical framework with findings of individual differences in the dopaminergic system during Pavlovian conditioning may explain why some individuals become addicted whereas others do not. Reinforcement learning models (Montague et al., 1996) give a powerful and deep account of the behavioural correlates of prediction error learning. Following McClure et al. (2003b), we have explained that this type of learning leads to representations in terms of model-free values, and that these capture key features of individual processing of motivational value, incentive salience assignment and sign-tracking. As such, it provides a framework within which neurobiology and behaviour relevant to addiction can be related in a computationally coherent manner (Redish et al., 2008; Dayan, 2009; Huys et al., 2013a), and forms one example of the application of computational neuroscience to psychiatric problems (Maia and Frank, 2011; Huys et al., 2011; Hasler, 2012; Montague et al., 2012; Huys et al., *subm*)

However, much remains to be done. While the description of model-free learning and the neurobiological details of the circuits computing prediction errors advance rapidly, our understanding of the representations and computations underlying model-based reasoning remain poorly defined. However, it is clear that addictions, and indeed many other affective psychiatric disorders, involve similar mechanisms.

6.5 FUNDING

We would like to acknowledge financial support by the National Institute of Health (1P01DA03165601) to SBF, the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) to QJMH (FOR 1617: grant RA1047/2-1) and the Swiss National Science Foundation to GH (32003B 138264) and PNT (PP00P1 128574 and CRSII3 141965).

6.6 ACKNOWLEDGEMENTS

We thank Peter Dayan, Maria Garbusow, Rike Petzschner and Terry Robinson for helpful comments and Katie Long for the drawings in Figure 3A,D.

REFERENCES

- Abler, B., Walter, H., Erk, S., Kammerer, H., and Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *Neuroimage*, 31(2):790–795.
- Balleine, B., Daw, N. D., and O’Doherty, J. (2009). Multiple forms of value learning and the function of dopamine. In Glimcher, P. W., Camerer, C., Fehr, E., and Poldrack, R., editors, *Neuroeconomics: Decision-making and the brain*.
- Barto, A., Sutton, R., and Anderson, C. (1983). Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13(5):834–846.
- Bayer, H. M. and Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–141.
- Bayer, H. M., Lau, B., and Glimcher, P. W. (2007). Statistics of midbrain dopamine neuron spike trains in the awake primate. *J Neurophysiol*, 98(3):1428–1439.
- Beck, A., Schlagenhauf, F., Wstenberg, T., Hein, J., Kienast, T., Kahnt, T., Schmack, K., Hgele, C., Knutson, B., Heinz, A., and Wrase, J. (2009). Ventral striatal activation during reward anticipation correlates with impulsivity in alcoholics. *Biol Psychiatry*.
- Beckmann, J. S., Marusich, J. A., Gipson, C. D., and Bardo, M. T. (2011). Novelty seeking, incentive salience and acquisition of cocaine self-administration in the rat. *Behav Brain Res*, 216(1):159–165.
- Beierholm, U., Guitart-Masip, M., Economides, M., Chowdhury, R., Dzel, E., Dolan, R., and Dayan, P. (2013). Dopamine modulates reward-related vigor. *Neuropsychopharmacology*, 38(8):1495–1503.
- Belin, D. and Everitt, B. J. (2008). Cocaine seeking habits depend upon dopamine-dependent serial connectivity linking the ventral with the dorsal striatum. *Neuron*, 57(3):432–441.
- Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press.
- Bello, E. P., Mateo, Y., Gelman, D. M., Noan, D., Shin, J. H., Low, M. J., Alvarez, V. A., Lovinger, D. M., and Rubinstein, M. (2011). Cocaine supersensitivity and enhanced motivation for reward in mice lacking dopamine d(2) autoreceptors. *Nat Neurosci*, 14(8):1033–1038.
- Berridge, K. C. (1996). Food reward: brain substrates of wanting and liking. *Neurosci Biobehav Rev*, 20(1):1–25.
- Berridge, K. C. (2004). Motivation concepts in behavioral neuroscience. *Physiol. Behav.*, 81:179–209.
- Berridge, K. C. (2007). The debate over dopamine’s role in reward: the case for incentive salience. *Psychopharmacology (Berl)*, 191(3):391–431.
- Berridge, K. C. (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. *Eur J Neurosci*, 35(7):1124–1143.
- Berridge, K. C. and Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res. Rev.*, 28(3):209–69.
- Berridge, K. C. and Robinson, T. E. (2003). Parsing reward. *Trends Neurosci*, 26(9):507–513.
- Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific.
- Boakes, R. (1977). Performance on learning to associate a stimulus with positive reinforcement. In Davis, H. and Hurwitz, H., editors, *Operant-Pavlovian Interactions*, pages 67–101. Lawrence Erlbaum Associates, Inc. Hillsdale, NJ.
- Boileau, I., Assaad, J.-M., Pihl, R. O., Benkelfat, C., Leyton, M., Diksic, M., Tremblay, R. E., and Dagher, A. (2003). Alcohol promotes dopamine release in the human nucleus accumbens. *Synapse*, 49(4):226–231.
- Boileau, I., Dagher, A., Leyton, M., Welfeld, K., Booij, L., Diksic, M., and Benkelfat, C. (2007). Conditioned dopamine release in humans: a positron emission tomography [11c]raclopride study with amphetamine. *J Neurosci*, 27(15):3998–4003.
- Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory*, 11(5):485–494.

- Bouton, M. E. (2006). *Learning and Behavior: A Contemporary Synthesis*. Sinauer, USA.
- Briand, L. A., Flagel, S. B., Garcia-Fuster, M. J., Watson, S. J., Akil, H., Sarter, M., and Robinson, T. E. (2008). Persistent alterations in cognitive function and prefrontal dopamine d2 receptors following extended, but not limited, access to self-administered cocaine. *Neuropsychopharmacology*, 33(12):2969–2980.
- Brischoux, F., Chakraborty, S., Brierley, D. I., and Ungless, M. A. (2009). Phasic excitation of dopamine neurons in ventral vta by noxious stimuli. *Proc Natl Acad Sci U S A*, 106(12):4894–4899.
- Bromberg-Martin, E. S., Matsumoto, M., Hong, S., and Hikosaka, O. (2010). A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J Neurophysiol*, 104(2):1068–1076.
- Buckholtz, J. W., Treadway, M. T., Cowan, R. L., Woodward, N. D., Li, R., Ansari, M. S., Baldwin, R. M., Schwartzman, A. N., Shelby, E. S., Smith, C. E., Kessler, R. M., and Zald, D. H. (2010). Dopaminergic network differences in human impulsivity. *Science*, 329(5991):532.
- Burke, C. J., Tobler, P. N., Baddeley, M., and Schultz, W. (2010). Neural mechanisms of observational learning. *Proc Natl Acad Sci U S A*, 107(32):14431–14436.
- Campbell, M., Hoane, A., et al. (2002). Deep Blue. *Artificial Intelligence*, 134(1-2):57–83.
- Cardinal, R. N., Parkinson, J. A., Lachenal, G., Halkerston, K. M., Rudarakanchana, N., Hall, J., Morrison, C. H., Howes, S. R., Robbins, T. W., and Everitt, B. J. (2002). Effects of selective excitotoxic lesions of the nucleus accumbens core, anterior cingulate cortex, and central nucleus of the amygdala on autoshaping performance in rats. *Behav Neurosci*, 116(4):553–567.
- Carter, B. L. and Tiffany, S. T. (1999). Meta-analysis of cue-reactivity in addiction research. *Addiction*, 94(3):327–340.
- Cavanagh, J., Eisenberg, I., Guitart-Masip, M., Huys, Q. J. M., and Frank, M. J. (2013). Frontal theta overrides pavlovian learning biases. *J. Neurosci.*, In Press:...
- Chen, B. T., Yau, H.-J., Hatch, C., Kusumoto-Yoshida, I., Cho, S. L., Hopf, F. W., and Bonci, A. (2013). Rescuing cocaine-induced prefrontal cortex hypoactivity prevents compulsive cocaine seeking. *Nature*, 496(7445):359–362.
- Clark, J. J., Collins, A. L., Sanford, C. A., and Phillips, P. E. M. (2013). Dopamine encoding of pavlovian incentive stimuli diminishes with extended training. *J Neurosci*, 33(8):3526–3532.
- Corbit, L. H. and Balleine, B. W. (2005). Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of Pavlovian-instrumental transfer. *J Neurosci*, 25(4):962–970.
- Corbit, L. H. and Balleine, B. W. (2011). The general and outcome-specific forms of pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. *J Neurosci*, 31(33):11786–11794.
- Courville, A. C., Daw, N., Gordon, G. J., and Touretzky, D. S. (2004). Model uncertainty in classical conditioning. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.
- Courville, A. C., Daw, N. D., and Touretzky, D. S. (2005). Similarity and discrimination in classical conditioning: A latent variable account. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 313–320. MIT Press, Cambridge, MA.
- Cox, S. M. L., Benkelfat, C., Dagher, A., Delaney, J. S., Durand, F., McKenzie, S. A., Kolivakis, T., Casey, K. F., and Leyton, M. (2009). Striatal dopamine responses to intranasal cocaine self-administration in humans. *Biol Psychiatry*, 65(10):846–850.
- Dagher, A. and Robbins, T. W. (2009). Personality, addiction, dopamine: insights from parkinson’s disease. *Neuron*, 61(4):502–510.
- Dalley, J. W. and Everitt, B. J. (2009). Dopamine receptors in the learning, memory and drug reward circuitry. *Semin Cell Dev Biol*, 20(4):403–410.
- Dalley, J. W. and Roiser, J. P. (2012). Dopamine, serotonin and impulsivity. *Neuroscience*, 215:42–58.
- D’Ardenne, K., McClure, S. M., Nystrom, L. E., and Cohen, J. D. (2008). Bold responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, 319(5867):1264–1267.
- Darke, S., Sims, J., McDonald, S., and Wickes, W. (2000). Cognitive impairment among methadone maintenance patients. *Addiction*, 95(5):687–695.
- Davey, G. C. and Cleland, G. G. (1982). Topography of signal-centered behavior in the rat: Effects of deprivation state and reinforcer type. *J Exp Anal Behav*, 38(3):291–304.
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans’ choices and striatal prediction errors. *Neuron*, 69(6):1204–1215.
- Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, 8(12):1704–1711.
- Daw, N. D. and Tobler, P. N. (2013). Value learning through reinforcement: the basics of dopamine and reinforcement learning. In Glimcher, P. W. and Fehr, E., editors, *Neuroeconomics*, pages 283–298. Academic Press, 2nd. edition.
- Day, J. J., Roitman, M. F., Wightman, R. M., and Carelli, R. M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat Neurosci*, 10(8):1020–1028.
- Dayan, P. (2009). Dopamine, reinforcement learning, and addiction. *Pharmacopsychiatry*, 42 Suppl 1:S56–

- S65.
- Dayan, P. and Berridge, K. C. (2013). Pavlovian values. *In prep*, 00:00.
- Dayan, P., Kakade, S., and Montague, P. R. (2000). Learning and selective attention. *Nat Neurosci*, 3 Suppl:1218–1223.
- Dayan, P. and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Curr Opin Neurobiol*, 18(2):185–196.
- de Lafuente, V. and Romo, R. (2011). Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions. *Proc Natl Acad Sci U S A*, 108(49):19767–19771.
- de Wit, H., Uhlenhuth, E. H., and Johanson, C. E. (1986). Individual differences in the reinforcing and subjective effects of amphetamine and diazepam. *Drug Alcohol Depend*, 16(4):341–360.
- de Wit, S., Corlett, P. R., Aitken, M. R., Dickinson, A., and Fletcher, P. C. (2009). Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans. *J Neurosci*, 29(36):11330–11338.
- DeJong, W. (1994). Relapse prevention: an emerging technology for promoting long-term drug abstinence. *Int. J. Addict.*, 29:681–705.
- Deroche-Gamonet, V., Belin, D., and Piazza, P. V. (2004). Evidence for addiction-like behavior in the rat. *Science*, 305(5686):1014–1017.
- Deserno, L., Beck, A., Huys, Q. J. M., Lorenz, R. C., Buchert, R., Buchholz, H. G., Plotkin, M., Kumakura, Y., Cumming, P., Heinze, H. J., Rapp, M. A., and Heinz, A. (2013). Chronic alcohol intake abolishes the relationship between dopamine synthesis capacity and learning signals in ventral striatum. *Under Review*, 00:00.
- Di Ciano, P., Cardinal, R. N., Cowell, R. A., Little, S. J., and Everitt, B. J. (2001). Differential involvement of nmda, ampa/kainate, and dopamine receptors in the nucleus accumbens core in the acquisition and performance of pavlovian approach behavior. *J Neurosci*, 21(23):9471–9477.
- Dickinson, A. and Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, 22(1):1–18.
- Dickinson, A. and Balleine, B. (2002). The role of learning in the operation of motivational systems. In Gallistel, R., editor, *Stevens' handbook of experimental psychology*, volume 3, pages 497–534. Wiley, New York.
- Dickinson, A. and Dearing, M. F. (1979). Appetitive-aversive interactions and inhibitory processes. In Dickinson, A. and Boakes, R. A., editors, *Mechanisms of learning and motivation*, pages 203–231. Erlbaum, Hillsdale, NJ.
- Dickinson, A., Smith, J., and Mirenowicz, J. (2000). Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. *Behav Neurosci*, 114(3):468–483.
- Dickinson, A., Wood, N., and Smith, J. W. (2002). Alcohol seeking by rats: action or habit? *Q J Exp Psychol B*, 55(4):331–348.
- Dreyer, J. K., Herrik, K. F., Berg, R. W., and Hounsgaard, J. D. (2010). Influence of phasic and tonic dopamine release on receptor activation. *J Neurosci*, 30(42):14273–14283.
- Düzel, E., Bunzeck, N., Guitart-Masip, M., Wittmann, B., Schott, B. H., and Tobler, P. N. (2009). Functional imaging of the human dopaminergic midbrain. *Trends Neurosci*, 32(6):321–328.
- Enomoto, K., Matsumoto, N., Nakai, S., Satoh, T., Sato, T. K., Ueda, Y., Inokawa, H., Haruno, M., and Kimura, M. (2011). Dopamine neurons learn to encode the long-term value of multiple future rewards. *Proc Natl Acad Sci U S A*, 108(37):15462–15467.
- Ersche, K. D., Bullmore, E. T., Craig, K. J., Shabbir, S. S., Abbott, S., Mller, U., Ooi, C., Suckling, J., Barnes, A., Sahakian, B. J., Merlo-Pich, E. V., and Robbins, T. W. (2010). Influence of compulsivity of drug abuse on dopaminergic modulation of attentional bias in stimulant dependence. *Arch Gen Psychiatry*, 67(6):632–644.
- Estes, W. and Skinner, B. (1941). Some quantitative aspects of anxiety. *J. Exp. Psychol*, 29:390–400.
- Evans, A. H., Pavese, N., Lawrence, A. D., Tai, Y. F., Appel, S., Doder, M., Brooks, D. J., Lees, A. J., and Piccini, P. (2006). Compulsive drug use linked to sensitized ventral striatal dopamine transmission. *Ann Neurol*, 59(5):852–858.
- Everitt, B. J., Dickinson, A., and Robbins, T. W. (2001). The neuropsychological basis of addictive behaviour. *Brain Res Brain Res Rev*, 36(2-3):129–138.
- Everitt, B. J. and Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat Neurosci*, 8(11):1481–1489.
- Fiorillo, C. D. (2013). Two dimensions of value: dopamine neurons represent reward but not aversiveness. *Science*, 341(6145):546–549.
- Fiorillo, C. D., Newsome, W. T., and Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. *Nat Neurosci*.
- Fiorillo, C. D., Song, M. R., and Yun, S. R. (2013a). Multiphasic temporal dynamics in responses of midbrain dopamine neurons to appetitive and aversive stimuli. *J Neurosci*, 33(11):4710–4725.
- Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by

- dopamine neurons. *Science*, 299(5614):1898–1902.
- Fiorillo, C. D., Yun, S. R., and Song, M. R. (2013b). Diversity and homogeneity in responses of midbrain dopamine neurons. *J Neurosci*, 33(11):4693–4709.
- Flagel, S. B., Akil, H., and Robinson, T. E. (2009). Individual differences in the attribution of incentive salience to reward-related cues: Implications for addiction. *Neuropharmacology*, 56 Suppl 1:139–148.
- Flagel, S. B., Cameron, C. M., Pickup, K. N., Watson, S. J., Akil, H., and Robinson, T. E. (2011a). A food predictive cue must be attributed with incentive salience for it to induce c-fos mRNA expression in corticostriatal-thalamic brain regions. *Neuroscience*, 196:80–96.
- Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., Akers, C. A., Clinton, S. M., Phillips, P. E. M., and Akil, H. (2011b). A selective role for dopamine in stimulus-reward learning. *Nature*, 469(7328):53–57.
- Flagel, S. B., Robinson, T. E., Clark, J. J., Clinton, S. M., Watson, S. J., Seeman, P., Phillips, P. E. M., and Akil, H. (2010). An animal model of genetic vulnerability to behavioral disinhibition and responsiveness to reward-related cues: implications for addiction. *Neuropsychopharmacology*, 35(2):388–400.
- Flagel, S. B., Waselus, M., Clinton, S. M., Watson, S. J., and Akil, H. (2013). Antecedents and consequences of drug abuse in rats selectively bred for high and low response to novelty. *Neuropharmacology*.
- Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *J Cogn Neurosci*, 17(1):51–72.
- Frank, M. J., Seeberger, L. C., and O'Reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in Parkinsonism. *Science*, 306(5703):1940–3.
- Franken, I. H. A., Hendriks, V. M., Stam, C. J., and Van den Brink, W. (2004). A role for dopamine in the processing of drug cues in heroin dependent patients. *Eur Neuropsychopharmacol*, 14(6):503–508.
- Freeman, T. P., Morgan, C. J. A., Beesley, T., and Curran, H. V. (2012). Drug cue induced overshadowing: selective disruption of natural reward processing by cigarette cues amongst abstinent but not satiated smokers. *Psychol Med*, 42(1):161–171.
- Ganesan, R. and Pearce, J. M. (1988). Effect of changing the unconditioned stimulus on appetitive blocking. *J Exp Psychol Anim Behav Process*, 14(3):280–291.
- Gelder, M., Harrison, P., and Cowen, P. (2006). *Shorter Oxford Textbook of Psychiatry*. Oxford University Press, Oxford, UK.
- Gershman, S. J., Blei, D. M., and Niv, Y. (2010). Context, learning, and extinction. *Psychol Rev*, 117(1):197–209.
- Gillan, C. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Voon, V., Apergis-Schoute, A. M., Fineberg, N. A., Sahakian, B. J., and Robbins, T. W. (2013). Enhanced avoidance habits in obsessive-compulsive disorder. *Biol Psychiatry*.
- Gillan, C. M., Pappmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., and de Wit, S. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *Am J Psychiatry*, 168(7):718–726.
- Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595.
- Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc Natl Acad Sci U S A*, 108 Suppl 3:15647–15654.
- Goldstein, R. Z., Leskovjan, A. C., Hoff, A. L., Hitzemann, R., Bashan, F., Khalsa, S. S., Wang, G.-J., Fowler, J. S., and Volkow, N. D. (2004). Severity of neuropsychological impairment in cocaine and alcohol addiction: association with metabolism in the prefrontal cortex. *Neuropsychologia*, 42(11):1447–1458.
- Grant, B. F., Stinson, F. S., Dawson, D. A., Chou, S. P., Dufour, M. C., Compton, W., Pickering, R. P., and Kaplan, K. (2004). Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: results from the national epidemiologic survey on alcohol and related conditions. *Arch Gen Psychiatry*, 61(8):807–816.
- Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annu Rev Neurosci*, 31:359–387.
- Grüsser, S. M., Wrase, J., Klein, S., Hermann, D., Smolka, M. N., Ruf, M., Weber-Fahr, W., Flor, H., Mann, K., Braus, D. F., and Heinz, A. (2004). Cue-induced activation of the striatum and medial prefrontal cortex is associated with subsequent relapse in abstinent alcoholics. *Psychopharmacology (Berl)*, 175(3):296–302.
- Guitart-Masip, M., Economides, M., Huys, Q. J. M., J, F. M., R, C., E, D., P, D., and J, D. R. (2013). Differential, but not opponent, effects of l-dopa and citalopram on action learning with reward and punishment. *Psychopharmacology*, In Press:.
- Guitart-Masip, M., Huys, Q. J. M., Fuentemilla, L., Dayan, P., Duzel, E., and Dolan, R. J. (2012). Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage*, 62(1):154–166.
- Hasler, G. (2012). Can the neuroeconomics revolution revolutionize psychiatry? *Neurosci Biobehav Rev*, 36(1):64–78.
- Hearst, E. and Jenkins, H. M. (1974). *Sign-tracking: The stimulus-reinforcer relation and directed action*. Psychonomic Society.
- Heinz, A., Beck, A., Wrase, J., Mohr, J., Obermayer, K., Gallinat, J., and Puls, I. (2009). Neurotransmitter

- systems in alcohol dependence. *Pharmacopsychiatry*, 42 Suppl 1:S95–S101.
- Heinz, A., Braus, D. F., Smolka, M. N., Wrase, J., Puls, I., Hermann, D., Klein, S., Grusser, S. S. M., Flor, H., Schumann, G., Mann, K., and Buchel, C. (2005). Amygdala-prefrontal coupling depends on a genetic variation of the serotonin transporter. *Nat Neurosci*, 8(1):20–1.
- Heinz, A., Dufeu, P., Kuhn, S., Dettling, M., Grf, K., Krten, I., Rommelspacher, H., and Schmidt, L. G. (1996). Psychopathological and behavioral correlates of dopaminergic sensitivity in alcohol-dependent patients. *Arch Gen Psychiatry*, 53(12):1123–1128.
- Heinz, A., Siessmeier, T., Wrase, J., Hermann, D., Klein, S., Grsser, S. M., Grsser-Sinopoli, S. M., Flor, H., Braus, D. F., Buchholz, H. G., Grnder, G., Schreckenberger, M., Smolka, M. N., Rsch, F., Mann, K., and Bartenstein, P. (2004). Correlation between dopamine d(2) receptors in the ventral striatum and central processing of alcohol cues and craving. *Am J Psychiatry*, 161(10):1783–1789.
- Hogarth, L., Balleine, B. W., Corbit, L. H., and Killcross, S. (2013). Associative learning mechanisms underpinning the transition from recreational drug use to addiction. *Ann N Y Acad Sci*, 1282:12–24.
- Hogarth, L. and Chase, H. W. (2011). Parallel goal-directed and habitual control of human drug-seeking: implications for dependence vulnerability. *J Exp Psychol Anim Behav Process*, 37(3):261–276.
- Hogarth, L., Dickinson, A., Wright, A., Kouvaraki, M., and Duka, T. (2007). The role of drug expectancy in the control of human drug seeking. *J Exp Psychol Anim Behav Process*, 33(4):484–496.
- Hollerman, J. R. and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat Neurosci*, 1(4):304–309.
- Holmes, N. M., Marchand, A. R., and Coutureau, E. (2010). Pavlovian to instrumental transfer: a neurobehavioural perspective. *Neurosci Biobehav Rev*, 34(8):1277–1295.
- Huys, Q. J. M., Beck, A., Dayan, P., and Heinz, A. (2013a). Neurobiological structure and computational understanding of addictive behaviour. In *Phenomenological neuropsychiatry: bridging the clinic and clinical neuroscience*. In Press.
- Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., and Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Comput Biol*, 7(4):e1002028.
- Huys, Q. J. M., Eshel, N., O’Nions, E., Sheridan, L., Dayan, P., and Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput Biol*, 8(3):e1002410.
- Huys, Q. J. M., Guitart-Masip, M., Dolan, R. J., and Dayan, P. (subm). Decision-theoretic psychiatry. *Clinical Psychological Science*, 00:00.
- Huys, Q. J. M., Pizzagalli, D. A., Bogdan, R., and Dayan, P. (2013b). Mapping anhedonia onto reinforcement learning: A behavioural meta-analysis. *Biol Mood Anxiety Disord*, 3(1):12.
- Ilango, A., Shumake, J., Wetzell, W., Scheich, H., and Ohl, F. W. (2012). The role of dopamine in the context of aversive stimuli with particular reference to acoustically signaled avoidance learning. *Front Neurosci*, 6:132.
- Iordanova, M. D., Westbrook, R. F., and Killcross, A. S. (2006). Dopamine activity in the nucleus accumbens modulates blocking in fear conditioning. *Eur J Neurosci*, 24(11):3265–3270.
- Jaffe, A., Gitisetan, S., Tarash, I., Pham, A., and Jentsch, J. (2010). Are nicotine-related cues susceptible to the blocking effect? In *SFN abstracts*.
- Johnson, A. and Redish, A. D. (2007). Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *J Neurosci*, 27(45):12176–12189.
- Johnson, P. M. and Kenny, P. J. (2010). Dopamine d2 receptors in addiction-like reward dysfunction and compulsive eating in obese rats. *Nat Neurosci*, 13(5):635–641.
- Jones, J. L., Esber, G. R., McDannald, M. A., Gruber, A. J., Hernandez, A., Mirens, A., and Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science*, 338(6109):953–956.
- Kahnt, T., Park, S. Q., Burke, C. J., and Tobler, P. N. (2012). How glitter relates to gold: similarity-dependent reward prediction errors in the human striatum. *J Neurosci*, 32(46):16521–16529.
- Kalivas, P. W. and Volkow, N. D. (2005). The neural basis of addiction: a pathology of motivation and choice. *Am J Psychiatry*, 162(8):1403–1413.
- Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In Campbell, B. A. and Church, R. M., editors, *Punishment and aversive behavior*. Appleton-Century-Crofts, New York.
- Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Reward-predicting activity of dopamine and caudate neurons—a possible mechanism of motivational control of saccadic eye movement. *J Neurophysiol*, 91(2):1013–1024.
- Keramati, M., Dezfouli, A., and Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput Biol*, 7(5):e1002055.
- Killcross, S. and Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb Cortex*, 13(4):400–408.
- Kim, K. M., Baratta, M. V., Yang, A., Lee, D., Boyden, E. S., and Fiorillo, C. D. (2012). Optogenetic mimicry

- of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. *PLoS One*, 7(4):e33612.
- Kirby, K. N., Petry, N. M., and Bickel, W. K. (1999). Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *J Exp Psychol Gen*, 128(1):78–87.
- Knutson, B. and Gibbs, S. E. B. (2007). Linking nucleus accumbens dopamine and blood oxygenation. *Psychopharmacology (Berl)*, 191(3):813–822.
- Kobayashi, S. and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. *J Neurosci*, 28(31):7837–7846.
- Koob, G. (1992). Dopamine, addiction and reward. In *Seminars in Neuroscience*, volume 4, pages 139–148. Elsevier.
- Kravitz, A. V., Tye, L. D., and Kreitzer, A. C. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nat Neurosci*, 15(6):816–818.
- Lesaint, F., Sigaud, O., Flagel, S., Robinson, T., and Khamassi, M. (2013). Modelling individual differences in the form of pavlovian conditioned approach responses: A dual learning systems approach with factored representations. *Under Review*, 00:00.
- Lex, A. and Hauber, W. (2008). Dopamine d1 and d2 receptors in the nucleus accumbens core and shell mediate pavlovian-instrumental transfer. *Learn Mem*, 15(7):483–491.
- Leyton, M., Boileau, I., Benkelfat, C., Diksic, M., Baker, G., and Dagher, A. (2002). Amphetamine-induced increases in extracellular dopamine, drug wanting, and novelty seeking: a pet/[11c]raclopride study in healthy men. *Neuropsychopharmacology*, 27(6):1027–1035.
- Lomanowska, A. M., Lovic, V., Rankine, M. J., Mooney, S. J., Robinson, T. E., and Kraemer, G. W. (2011). Inadequate early social experience increases the incentive salience of reward-related cues in adulthood. *Behav Brain Res*, 220(1):91–99.
- Lovibond, P. F. (1983). Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. *J Exp Psychol Anim Behav Process*, 9(3):225–247.
- Lovic, V., Saunders, B. T., Yager, L. M., and Robinson, T. E. (2011). Rats prone to attribute incentive salience to reward cues are also prone to impulsive action. *Behav Brain Res*, 223(2):255–261.
- Lucantonio, F., Stalnaker, T. A., Shaham, Y., Niv, Y., and Schoenbaum, G. (2012). The impact of orbitofrontal dysfunction on cocaine addiction. *Nat Neurosci*, 15(3):358–366.
- Mahler, S. V. and de Wit, H. (2010). Cue-reactors: individual differences in cue-induced craving after food or smoking abstinence. *PLoS One*, 5(11):e15475.
- Maia, T. V. and Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nat Neurosci*, 14(2):154–162.
- Martinez, D., Gil, R., Slifstein, M., Hwang, D.-R., Huang, Y., Perez, A., Kegeles, L., Talbot, P., Evans, S., Krystal, J., Laruelle, M., and Abi-Dargham, A. (2005). Alcohol dependence is associated with blunted dopamine transmission in the ventral striatum. *Biol Psychiatry*, 58(10):779–786.
- Martinez, D., Greene, K., Broft, A., Kumar, D., Liu, F., Narendran, R., Slifstein, M., Van Heertum, R., and Kleber, H. D. (2009). Lower level of endogenous dopamine in patients with cocaine dependence: findings from pet imaging of d(2)/d(3) receptors following acute dopamine depletion. *Am J Psychiatry*, 166(10):1170–1177.
- Matsumoto, M. and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, 459(7248):837–841.
- Mazzoni, P., Hristova, A., and Krakauer, J. W. (2007). Why don't we move faster? Parkinson's disease, movement vigor, and implicit motivation. *J Neurosci*, 27(27):7105–7116.
- McClure, S. M., Berns, G. S., and Montague, P. R. (2003a). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, 38(2):339–46.
- McClure, S. M., Daw, N. D., and Montague, P. R. (2003b). A computational substrate for incentive salience. *TINS*, 26:423–8.
- McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., and Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *J Neurosci*, 31(7):2700–2705.
- Meyer, P. J., Lovic, V., Saunders, B. T., Yager, L. M., Flagel, S. B., Morrow, J. D., and Robinson, T. E. (2012). Quantifying individual variation in the propensity to attribute incentive salience to reward cues. *PLoS One*, 7(6):e38987.
- Mileykovskiy, B. and Morales, M. (2011). Duration of inhibition of ventral tegmental area dopamine neurons encodes a level of conditioned fear. *J Neurosci*, 31(20):7471–7476.
- Milton, A. L. and Everitt, B. J. (2010). The psychological and neurochemical mechanisms of drug memory reconsolidation: implications for the treatment of addiction. *Eur J Neurosci*, 31(12):2308–2319.
- Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J. Neurosci.*, 16(5):1936–47.
- Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends Cogn Sci*, 16(1):72–80.

- Morgan, D., Grant, K. A., Gage, H. D., Mach, R. H., Kaplan, J. R., Prioleau, O., Nader, S. H., Buchheimer, N., Ehrenkaufer, R. L., and Nader, M. A. (2002). Social dominance in monkeys: dopamine d2 receptors and cocaine self-administration. *Nat Neurosci*, 5(2):169–174.
- Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nat Neurosci*, 9(8):1057–1063.
- Müller, C. P. and Schumann, G. (2011). To use or not to use: Expanding the view on non-addictive psychoactive drug consumption and its implications. *Behav Brain Sci*, 34(6):328–347.
- Murschall, A. and Hauber, W. (2006). Inactivation of the ventral tegmental area abolished the general excitatory influence of pavlovian cues on instrumental performance. *Learn Mem*, 13(2):123–126.
- Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron*, 41(2):269–280.
- Nelson, A. and Killcross, S. (2006). Amphetamine exposure enhances habit formation. *J Neurosci*, 26(14):3805–3812.
- Nelson, A. J. D. and Killcross, S. (2013). Accelerated habit formation following amphetamine exposure is reversed by d1, but enhanced by d2, receptor antagonists. *Front Neurosci*, 7:76.
- Niv, Y., Daw, N. D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology (Berl)*, 191(3):507–520.
- Nomoto, K., Schultz, W., Watanabe, T., and Sakagami, M. (2010). Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. *J Neurosci*, 30(32):10692–10702.
- O'Brien, M. S. and Anthony, J. C. (2005). Risk of becoming cocaine dependent: epidemiological estimates for the united states, 2000-2001. *Neuropsychopharmacology*, 30(5):1006–1018.
- O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669):452–4.
- O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–337.
- Olds, J. (1956). A preliminary mapping of electrical reinforcing effects in the rat brain. *J Comp Physiol Psychol*, 49(3):281–285.
- Ostlund, S. B. and Balleine, B. W. (2008). On habits and addiction: An associative analysis of compulsive drug seeking. *Drug Discov Today Dis Models*, 5(4):235–245.
- Ostlund, S. B., Maidment, N. T., and Balleine, B. W. (2010). Alcohol-paired contextual cues produce an immediate and selective loss of goal-directed action in rats. *Front Integr Neurosci*, 4.
- Otto, A. R., Gershman, S. J., Markman, A. B., and Daw, N. D. (2013). The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol Sci*, 24(5):751–761.
- Oyama, K., Herndi, I., Iijima, T., and Tsutsui, K.-I. (2010). Reward prediction error coding in dorsal striatal neurons. *J Neurosci*, 30(34):11447–11457.
- Panlilio, L. V., Thorndike, E. B., and Schindler, C. W. (2007). Blocking of conditioning to a cocaine-paired stimulus: testing the hypothesis that cocaine perpetually produces a signal of larger-than-expected reward. *Pharmacol Biochem Behav*, 86(4):774–777.
- Parker, J. G., Zweifel, L. S., Clark, J. J., Evans, S. B., Phillips, P. E. M., and Palmiter, R. D. (2010). Absence of nmda receptors in dopamine neurons attenuates dopamine release but not conditioned approach during pavlovian conditioning. *Proc Natl Acad Sci U S A*, 107(30):13491–13496.
- Parkinson, J. A., Dalley, J. W., Cardinal, R. N., Bamford, A., Fehner, B., Lachenal, G., Rudarakanchana, N., Halkerston, K. M., Robbins, T. W., and Everitt, B. J. (2002). Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive pavlovian approach behaviour: implications for mesoaccumbens dopamine function. *Behav Brain Res*, 137(1-2):149–163.
- Pearce, J. M. (1997). *Animal Learning and Cognition*. Psychology Press, Hove, UK.
- Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., and Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106):1042–1045.
- Pezzulo, G., Rigoli, F., and Chersi, F. (2013). The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Front Psychol*, 4:92.
- Pfeiffer, B. E. and Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74–79.
- Prévost, C., Liljeholm, M., Tyszka, J. M., and O'Doherty, J. P. (2012). Neural correlates of specific and general pavlovian-to-instrumental transfer within human amygdalar subregions: a high-resolution fmri study. *J Neurosci*, 32(24):8383–8390.
- Prévost, C., McNamee, D., Jessup, R. K., Bossaerts, P., and O'Doherty, J. P. (2013). Evidence for model-based computations in the human amygdala during pavlovian conditioning. *PLoS Comput Biol*, 9(2):e1002918.
- Redish, A. D. (2004). Addiction as a computational process gone awry. *Science*, 306(5703):1944–1947.
- Redish, A. D., Jensen, S., and Johnson, A. (2008). A unified framework for addiction: vulnerabilities in the decision process. *Behav Brain Sci*, 31(4):415–37; discussion 437–87.
- Rescorla, R. and Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, pages 64–99.

- Rescorla, R. A. and Solomon, R. L. (1967). Two-process learning theory: Relationships between pavlovian conditioning and instrumental learning. *Psychol Rev*, 74(3):151–182.
- Richard, J. M., Plawecki, A. M., and Berridge, K. C. (2013). Nucleus accumbens gabaergic inhibition generates intense eating and fear that resists environmental retuning and needs no local dopamine. *Eur J Neurosci*, 37(11):1789–1802.
- Robbins, T. W., Gillan, C. M., Smith, D. G., de Wit, S., and Ersche, K. D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry. *Trends Cogn Sci*, 16(1):81–91.
- Robinson, M. J. F. and Berridge, K. C. (2013). Instant transformation of learned repulsion into motivational “wanting”. *Curr Biol*, 23(4):282–289.
- Robinson, T. E. and Flagel, S. B. (2009). Dissociating the predictive and incentive motivational properties of reward-related cues through the study of individual differences. *Biol Psychiatry*, 65(10):869–873.
- Roesch, M. R., Calu, D. J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat Neurosci*, 10(12):1615–1624.
- Rossi, M. A., Sukharnikova, T., Hayrapetyan, V. Y., Yang, L., and Yin, H. H. (2013). Operant self-stimulation of dopamine neurons in the substantia nigra. *PLoS One*, 8(6):e65799.
- Russell, S. J. and Wefald, E. H. (1991). *Do the right thing: studies in limited rationality*. MIT press.
- Rutledge, R. B., Dean, M., Caplin, A., and Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *J Neurosci*, 30(40):13525–13536.
- Salamone, J. D., Correa, M., Farrar, A. M., Nunes, E. J., and Pardo, M. (2009). Dopamine, behavioral economics, and effort. *Front Behav Neurosci*, 3:13.
- Satoh, T., Nakai, S., Sato, T., and Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. *J Neurosci*, 23(30):9913–9923.
- Saunders, B. T. and Robinson, T. E. (2010). A cocaine cue acts as an incentive stimulus in some but not others: implications for addiction. *Biol Psychiatry*, 67(8):730–736.
- Saunders, B. T. and Robinson, T. E. (2011). Individual variation in the motivational properties of cocaine. *Neuropsychopharmacology*, 36(8):1668–1676.
- Saunders, B. T. and Robinson, T. E. (2012). The role of dopamine in the accumbens core in the expression of pavlovian-conditioned responses. *Eur J Neurosci*, 36(4):2521–2532.
- Saunders, B. T. and Robinson, T. E. (2013). Individual variation in resisting temptation: Implications for addiction. *Neurosci Biobehav Rev*, 37(9PA):1955–1975.
- Saunders, B. T., Yager, L. M., and Robinson, T. E. (2013a). Cue-evoked cocaine “craving”: role of dopamine in the accumbens core. *J Neurosci*, 33(35):13989–14000.
- Saunders, B. T., Yager, L. M., and Robinson, T. E. (2013b). Preclinical studies shed light on individual variation in addiction vulnerability. *Neuropsychopharmacology*, 38(1):249–250.
- Schlagenhauf, F., Rapp, M. A., Huys, Q. J. M., Beck, A., Wstenberg, T., Deserno, L., Buchholz, H.-G., Kalbitzer, J., Buchert, R., Bauer, M., Kienast, T., Cumming, P., Plotkin, M., Kumakura, Y., Grace, A. A., Dolan, R. J., and Heinz, A. (2013). Ventral striatal prediction error signaling is associated with dopamine synthesis capacity and fluid intelligence. *Hum Brain Mapp*, 34(6):1490–1499.
- Schoenbaum, G., Roesch, M. R., Stalnaker, T. A., and Takahashi, Y. K. (2009). A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nat Rev Neurosci*, 10(12):885–892.
- Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J Neurophysiol*, 80(1):1–27.
- Schultz, W. (2013). Updating dopamine reward signals. *Curr Opin Neurobiol*, 23(2):229–238.
- Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of neuroscience*, 13(3):900.
- Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.
- Schwabe, L. and Wolf, O. T. (2009). Stress prompts habit behavior in humans. *J Neurosci*, 29(22):7191–7198.
- Sebold, M., Deserno, L., Nebe, S., Schad, D., Garbusow, M., Hägele, C., Keller, J., Jünger, E., Kathmann, N., Smolka, M., Rapp, M. A., Schlagenhauf, F., Heinz, A., and M, H. Q. J. (subm). Model-based and model-free decisions in alcohol dependence. *Neuropsychobiology*, 00:00.
- Seligman, M. E. (1970). On the generality of the laws of learning. *Psychol Review*, 77(5):406–18.
- Seymour, B., O’Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., and Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, 429(6992):664–667.
- Shallice, T. (1982). Specific impairments of planning. *Philos Trans R Soc Lond B Biol Sci*, 298(1089):199–209.
- Shiflett, M. W. and Balleine, B. W. (2010). At the limbic-motor interface: disconnection of basolateral amygdala from nucleus accumbens core and shell reveals dissociable components of incentive motivation. *Eur J Neurosci*, 32(10):1735–1743.
- Shiner, T., Seymour, B., Wunderlich, K., Hill, C., Bhatia, K. P., Dayan, P., and Dolan, R. J. (2012). Dopamine and performance in a reinforcement learning task: evidence from parkinson’s disease. *Brain*, 135(Pt 6):1871–1883.

- Simon, D. A. and Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision task in humans. *J Neurosci*, 31(14):5526–5539.
- Smith, K. S. and Graybiel, A. M. (2013). A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron*, 79(2):361–374.
- Smith, K. S., Virkud, A., Deisseroth, K., and Graybiel, A. M. (2012). Reversible online control of habitual behavior by optogenetic perturbation of medial prefrontal cortex. *Proc Natl Acad Sci U S A*, 109(46):18932–18937.
- Spicer, J., Galvan, A., Hare, T. A., Voss, H., Glover, G., and Casey, B. (2007). Sensitivity of the nucleus accumbens to violations in expectation of reward. *Neuroimage*, 34(1):455–461.
- Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., and Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nat Neurosci*, 16(7):966–973.
- Sulzer, D. (2011). How addictive drugs disrupt presynaptic dopamine neurotransmission. *Neuron*, 69(4):628–649.
- Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
- Takahashi, Y. K., Roesch, M. R., Wilson, R. C., Toreson, K., O'Donnell, P., Niv, Y., and Schoenbaum, G. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nat Neurosci*.
- Takikawa, Y., Kawagoe, R., and Hikosaka, O. (2004). A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. *J Neurophysiol*, 92(4):2520–2529.
- Talmi, D., Seymour, B., Dayan, P., and Dolan, R. J. (2008). Human Pavlovian-instrumental transfer. *J Neurosci*, 28(2):360–368.
- Tan, K. R., Yvon, C., Turiault, M., Mirzabekov, J. J., Doehner, J., Laboube, G., Deisseroth, K., Tye, K. M., and Lscher, C. (2012). Gaba neurons of the vta drive conditioned place aversion. *Neuron*, 73(6):1173–1183.
- Thanos, P. K., Volkow, N. D., Freimuth, P., Umegaki, H., Ikari, H., Roth, G., Ingram, D. K., and Hitzemann, R. (2001). Overexpression of dopamine d2 receptors reduces alcohol self-administration. *J Neurochem*, 78(5):1094–1103.
- Tobler, P. N., Dickinson, A., and Schultz, W. (2003). Coding of predicted reward omission by dopamine neurons in a conditioned inhibition paradigm. *J Neurosci*, 23(32):10402–10410.
- Tobler, P. N., Fiorillo, C. D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, 307(5715):1642–1645.
- Tobler, P. N., O'Doherty, J. P., Dolan, R. J., and Schultz, W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. *J Neurophysiol*, 95(1):301–310.
- Tobler, P. N., O'Doherty, J. P., Dolan, R. J., and Schultz, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *J Neurophysiol*, 97(2):1621–1632.
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, 55(4):189.
- Tom, S. M., Fox, C. R., Trepel, C., and Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518.
- Tomie, A., Aguado, A. S., Pohorecky, L. A., and Benjamin, D. (1998). Ethanol induces impulsive-like responding in a delay-of-reward operant choice procedure: impulsivity predicts autoshaping. *Psychopharmacology (Berl)*, 139(4):376–382.
- Tricomi, E., Balleine, B. W., and O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *Eur J Neurosci*, 29(11):2225–2232.
- Tsai, H.-C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., de Lecea, L., and Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science*, 324(5930):1080–1084.
- Valentin, V. V., Dickinson, A., and O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *J Neurosci*, 27(15):4019–4026.
- van der Meer, M. A. A. and Redish, A. D. (2009). Covert expectation-of-reward in rat ventral striatum at decision points. *Front Integr Neurosci*, 3:1.
- Vanderschuren, L. J. M. J. and Everitt, B. J. (2004). Drug seeking becomes compulsive after prolonged cocaine self-administration. *Science*, 305(5686):1017–1019.
- Volkow, N. D., Fowler, J. S., Wang, G. J., Baler, R., and Telang, F. (2009). Imaging dopamine's role in drug abuse and addiction. *Neuropharmacology*, 56 Suppl 1:3–8.
- Volkow, N. D., Fowler, J. S., Wang, G.-J., and Goldstein, R. Z. (2002). Role of dopamine, the frontal cortex and memory circuits in drug addiction: insight from imaging studies. *Neurobiol Learn Mem*, 78(3):610–624.
- Volkow, N. D., Wang, G.-J., Telang, F., Fowler, J. S., Logan, J., Childress, A.-R., Jayne, M., Ma, Y., and Wong, C. (2006). Cocaine cues and dopamine in dorsal striatum: mechanism of craving in cocaine addiction. *J Neurosci*, 26(24):6583–6588.
- Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, 412(6842):43–48.
- Walton, M. E., Behrens, T. E. J., Buckley, M. J., Rudebeck, P. H., and Rushworth, M. F. S. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron*, 65(6):927–939.
- Wassum, K. M., Ostlund, S. B., Balleine, B. W., and Maidment, N. T. (2011). Differential dependence of

- pavlovian incentive motivation and instrumental incentive learning processes on dopamine signaling. *Learn Mem*, 18(7):475–483.
- Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., and Lindenmeyer, J. (2011). Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychol Sci*, 22(4):490–497.
- Willuhn, I., Burgeno, L. M., Everitt, B. J., and Phillips, P. E. M. (2012). Hierarchical recruitment of phasic dopamine signaling in the striatum during the progression of cocaine use. *Proc Natl Acad Sci U S A*, 109(50):20703–20708.
- Wilson, S. J., Sayette, M. A., and Fiez, J. A. (2004). Prefrontal responses to drug cues: a neurocognitive analysis. *Nat Neurosci*, 7(3):211–214.
- Witten, I. B., Steinberg, E. E., Lee, S. Y., Davidson, T. J., Zalocusky, K. A., Brodsky, M., Yizhar, O., Cho, S. L., Gong, S., Ramakrishnan, C., Stuber, G. D., Tye, K. M., Janak, P. H., and Deisseroth, K. (2011). Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. *Neuron*, 72(5):721–733.
- Wrase, J., Schlagenhauf, F., Kienast, T., Wstenberg, T., Bermpohl, F., Kahnt, T., Beck, A., Strhle, A., Juckel, G., Knutson, B., and Heinz, A. (2007). Dysfunction of reward processing correlates with alcohol craving in detoxified alcoholics. *Neuroimage*, 35(2):787–794.
- Wunderlich, K., Dayan, P., and Dolan, R. J. (2012a). Mapping value based planning and extensively trained choice in the human brain. *Nat Neurosci*, 15(5):786–791.
- Wunderlich, K., Smittenaar, P., and Dolan, R. J. (2012b). Dopamine enhances model-based over model-free choice behavior. *Neuron*, 75(3):418–424.
- Wyvell, C. L. and Berridge, K. C. (2000). Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: enhancement of reward "wanting" without enhanced "liking" or response reinforcement. *J Neurosci*, 20(21):8122–8130.
- Yin, H. H., Knowlton, B. J., and Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.*, 19(1):181–9.
- Yin, H. H., Ostlund, S. B., and Balleine, B. W. (2008). Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *Eur J Neurosci*, 28(8):1437–1448.
- Yin, H. H., Ostlund, S. B., Knowlton, B. J., and Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *Eur J Neurosci*, 22(2):513–523.
- Zaghloul, K. A., Blanco, J. A., Weidemann, C. T., McGill, K., Jaggi, J. L., Baltuch, G. H., and Kahana, M. J. (2009). Human substantia nigra neurons encode unexpected financial rewards. *Science*, 323(5920):1496–1499.