Published in final edited form as:
Cogn Affect Behav Neurosci. 2014 June ; 14(2): 473–492. doi:10.3758/s13415-014-0277-8.

Model-Based and Model-Free Pavlovian Reward Learning: Revaluation, Revision and Revelation

Peter Dayan[1,*] and **Kent C. Berridge**[2,*]

[1]Gatsby Computational Neuroscience Unit, University College London

[2]Department of Psychology, University of Michigan

Abstract

Evidence supports at least two methods for learning about reward and punishment and making predictions for guiding actions. One method, called model-free, progressively acquires cached estimates of the long-run values of circumstances and actions from retrospective experience. The other method, called model-based, uses representations of the environment, expectations and prospective calculations to make cognitive predictions of future value. Extensive attention has been paid to both methods in computational analyses of instrumental learning. By contrast, although a full computational analysis has been lacking, Pavlovian learning and prediction has typically been presumed to be solely model-free. Here, we revise that presumption and review compelling evidence from Pavlovian revaluation experiments showing that Pavlovian predictions can involve their own form of model-based evaluation. In model-based Pavlovian evaluation, prevailing states of the body and brain influence value computations, and thereby produce powerful incentive motivations that can sometimes be quite new. We consider the consequences of this revised Pavlovian view for the computational landscape of prediction, response and choice. We also revisit differences between Pavlovian and instrumental learning in the control of incentive motivation.

1) Introduction

Pavlovian cues often elicit motivations to pursue and consume the rewards (or avoid the threats) with which they have been associated. The cues are called conditioned stimuli or CSs; the rewards or threats are called unconditioned stimuli or UCSs. For addicts and sufferers from related compulsive urges, cue-triggered motivations may become quite powerful and maladaptive; they also underpin various lucrative industries (Bushong, King, Camerer, & Rangel, 2010). Pavlovian learning and responding interacts in a rich and complex manner with instrumental learning and responding, in which subjects make choices contingent on expectations or past experience of the outcomes to which they lead.

Computational analyses of instrumental learning (involved in predicting which actions will be rewarded) have paid substantial attention to the critical distinction between *model-free* and *model-based* forms of learning and computation (Figure 1). Model-based strategies generate goal-directed choices employing a model or cognitive-style representation, which is

[*]Corresponding authors: dayan@gatsby.ucl.ac.ukberridge@umich.edu.

an internal map of events and stimuli from the external world (Daw, Niv, & Dayan, 2005; Dickinson & Balleine, 2002; Doya, 1999). That internal model supports prospective assessment of the consequences of taking particular actions. By contrast, model-free strategies have no model of outside events, but instead merely learn by caching information about the utilities of outcomes encountered on past interactions with the environment. This generates direct rules for how to behave, or propensities for performing particular actions, based on predictions of the long-run values of actions. Model-free values can be described as being free-floating, since they can become detached from any specific outcome. The model-based/model-free distinction has been experimentally highly fruitful (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Fermin, Yoshida, Ito, Yoshimoto, & Doya, 2010; Gläscher, Daw, Dayan, & O'Doherty, 2010; Wunderlich, Dayan, & Dolan, 2012). For example, model-based mechanisms are held to held to produce cognitive or flexibly goal-directed instrumental behaviour, whereas model-free mechanisms have often been treated as producing automatic instrumental stimulus-response habits (Daw et al, 2005) (though cf. (Dezfouli & Balleine, 2013)). There are also intermediate points between model-based and model-free instrumental control that we briefly discuss below.

What makes learning Pavlovian is that the conditioned response is directly elicited by a CS that is predictive of a UCS, without regard to the effect of the response on the provision or omission of that UCS (Mackintosh, 1983). This offers the significant efficiency advantage of substituting genotypic for phenotypic search amongst a potentially huge range of possible actions for one that is usually appropriate to a circumstance, but at the expense of inflexibility of response in particular cases. By contrast with instrumental learning, computational analyses of Pavlovian learning have, with only few exceptions (Doll, Simon, & Daw, 2012), presumed the computation of prediction to be model-free, leading to simple stored caches of stimulus-value associations. However, here we conduct a closer inspection of model-free and model-based alternatives specifically for Pavlovian learning and value predictions, attempting to meld recent insights from affective neuroscience studies of incentive motivation. We will conclude that model-based computations can play a critical role in Pavlovian learning and motivation, and this generates flexibility in at least some affective/motivational responses to a CS (Figure 1).

In order to illuminate the contrast between model-free and model-based predictions in Pavlovian situations, we draw on an illustrative experiment (which we call the 'Dead Sea salt' experiment) recently performed with rats by M.J. F. Robinson & Berridge (Robinson & Berridge, 2013). This experiment built on many past demonstrations that inducing a novel body need (sodium appetite) can reveal latent learning about salty food sources. Those sources may not have been attractive during learning, but can be used adaptively when a sodium need state is induced at a later time. For example, rats suddenly put into a salt appetite state will appropriately return to, work for, or even ingest, cues previously associated with salt that had no particular value to them when learned about earlier (Balleine, 1994; Dickinson, 1986; Fudim, 1978; Krieckhaus & Wolf, 1968; Rescorla & Freberg, 1978; Schulkin, Arnell, & Stellar, 1985; Stouffer & White, 2005; Wirsig & Grill, 1982). Sodium need also allows Pavlovian CS stimuli related to salt directly to undergo an hedonic transformation to become 'liked' when re-encountered in a relevant appetite state (i.e., CS alliesthesia, similar to alliesthesia of the salty UCS) (Berridge & Schulkin, 1989).

In the Dead Sea salt experiment (depicted in figure 2), a distinctive Pavlovian CS (the insertion of a lever through a wall into the chamber accompanied by a sound) was first paired with an inescapable disgusting UCS (Robinson & Berridge, 2013). The disgusting UCS was an intra-oral squirt of a saline solution whose high sodium chloride concentration, equivalent to that in the Dead Sea (i.e., triple that of ordinary seawater), made it very aversive. Simultaneously, a different CS (a lever inserted from the opposite side of chamber accompanied by a different sound) predicted a different pleasant UCS squirt of sweet sucrose solution into the mouth. The rats approached and nibbled the sucrose-related CS lever, but duly learned to be spatially repulsed by the salt-related CS whenever the lever appeared, physically "turning away and sometimes pressing themselves against the opposite wall" (p.283) as though trying to escape from the repulsive CS lever and keep as far away as physically possible. This is a prime case of appetitive versus aversive Pavlovian conditioning (Rescorla, 1988), with the escape response being drawn from a species-typical defensive repertoire and the appetitive response drawn from an ingestive repertoire. Such Pavlovian responses are elicited by cues that predict their respectively valenced outcomes, albeit somewhat adapted to the nature both of the CS and the UCS. The Pavlovian responses here achieved no instrumental benefit, and would likely have persisted even if this had actually decreased sucrose delivery or increased the probability of noxious salt delivery (as in (Anson, Bender, & Melvin, 1969; Fowler & Miller, 1963; Morse, Mead, & Kelleher, 1967)).

On a subsequent day, the rats were injected for the first time ever with the drugs deoxycorticosterone and furosemide. These mimic brain signals normally triggered by angiotensin II and aldosterone hormones under a state of salt deprivation (which the rats had never previously experienced). In their new state of salt appetite, the rats were then again presented with the lever CS, but in extinction (i.e., without the provision of any outcome). Their Pavlovian behaviour toward the CS in the new appetite state was reassessed, as was the activation of an immediate early gene in neurons as a signature of neural activity (cFos gene translation into Fos protein).

In the new condition, far from eliciting repulsion as before, the salty-related CS lever suddenly and specifically now became nearly as strongly attractive as the sweet-related lever (appetitive engagement with salt-associated CS increased by a factor of more than 10 compared with the pre-deprivation training days), so that the metal CS object was avidly approached, sniffed, grasped and nibbled (Robinson & Berridge, 2013). These novel salt-related CS responses were again Pavlovian, achieving no instrumental benefit (the metal lever was not salty, and pressing it had never obtained salt). The transformation of the motivation (creating what is known as a motivational magnet) occurred on the very first presentations of the CS in the new state, before the newly positive valence of the salty UCS taste had been experienced, and so without any new learning about its altered UCS or new CS-UCS pairing (Figure 2). No change in behaviour was seen toward the sucrose-associated lever, nor toward a third, control, lever that predicted nothing and typically was behaviourally ignored. Sometimes the salt-associated CS also elicited affective orofacial 'liking' reactions in the new appetite state that would ordinarily be elicited by a palatable taste UCS, such as licking of the lips or paws (though the rats had never yet tasted the concentrated NaCl as positively ‘liked’ in their new appetite state) (Robinson & Berridge, 2013).

These behavioural changes consequent on first reencountering the salt-related CS lever in the new appetite state were not the only new observation. Neurobiologically, activity in a collection of mesocorticolimbic brain areas was also dramatically upregulated by the combination of a) reencountering the CS+ lever simultaneously with b) being in the new appetite state (Figure 2). Fos was elevated in the core and rostral shell of the nucleus accumbens, as well as in the ventral tegmental area (VTA) and the rostral ventral pallidum, and in infralimbic and orbitofrontal regions of the prefrontal cortex (Robinson & Berridge, 2013). At least some of those brain areas, and particularly the neuromodulator dopamine (projected from the VTA to the nucleus accumbens and other structures), play a key role in the motivational attribution of incentive value to Pavlovian stimuli, a process known as incentive salience, which makes attributed stimuli (e.g., CSs as well as UCSs) become positively ‘wanted’. The changes in mesolimbic structures were not merely a function of increased physiological drive, but rather also required the CS+ in the new state. No significant Fos elevation at all was detected in ventral tegmentum given just the isolated state of salt appetite by itself in absence of CS+ lever, and only one-third as high or less Fos elevation was seen in nucleus accumbens regions as when salt CS and appetite state were combined together. This apparent requirement for simultaneous CS plus appetite state to activate mesolimbic circuits maximally, replicates a previous finding that firing of neurons in ventral pallidum was also elevated only by simultaneous salt CS plus appetite (but again without actually tasting the NaCl in the deprived state) (Tindell, Smith, Berridge, & Aldridge, 2009), compared to either the deprived state alone or to CS encounters alone in the normal state. The earlier study also used a diffuse auditory CS that could not be spatially approached, ensuring that CS value and not elicited appetitive behaviour was driving the neural activation (Tindell et al., 2009).

These new experiments helped resolve an important motivational question as to whether sudden appetitive behaviour toward the saltiness source is motivated simply to alleviate negative distress of the salt appetite state (i.e., reduce aversive drive), or whether Pavlovian CSs for saltiness actually become positively ‘wanted’, endowed with incentive salience when re-encountered in a novel relevant state. Pavlovian CSs that are the targets of incentive salience capture attention, are attractive, stimulate approach, and even elicit some forms of consumption behaviour, almost as if they had come to share some key characteristics with the food, drug, or other reward UCSs themselves (Berridge, 2007; Toates, 1986). We interpret the results of the Dead Sea salt experiment as demonstrating spontaneous generation of positive Pavlovian incentive salience in a fashion we suggest is model-based. It also shows that the CS’s transformation of Pavlovian motivation can be so powerful as to reverse nearly instantly an intense earlier learned repulsion into a suddenly positive, intense incentive ‘want’.

We focus on the Pavlovian reversal from repulsion to attraction because it is an especially vivid example of state-induced transformation in CS value. However, it is only one exemplar of a wider and long-studied class of revaluation changes in Pavlovian responses that we suggest demand explanation in terms of similar model-based mechanisms, involving stimulus-stimulus associations that preserve the details about the identities of events that have been learned (Dickinson, 1986; Holland, 1990; Holland, Lasseter, & Agarwal, 2008) (Bouton & Moody, 2004; Holland et al., 2008; Rescorla, 1988; Rizley & Rescorla, 1972; Zener & McCurdy, 1939). In all of these, individuals show that they can use learned information about the identity of a UCS that is associated with a particular CS when new information is later added to that CS (for example, developing a taste aversion to an absent UCS, when its associated CS later becomes paired associatively with illness (Holland, 1990). Other related cases indicate that the incentive salience of CSs can be similarly multiplied at the time of CS re-encounter as a result of neurobiological activations of mesolimbic systems, induced either by sudden dopamine/opioid pharmacological stimulation or by drug-induced neural sensitization interposed between CS-UCS training and CS reencounter (Difeliceantonio & Berridge, 2012; Smith, Berridge, & Aldridge, 2011; Pecina & Berridge, 2013; Vezina & Leyton, 2009; Wyvell & Berridge, 2001; Tindell et al., 2005).

What is needed is a computational dissection of the way that such Pavlovian transformations in CS-triggered motivation happen. We seek the same quality of understanding for Pavlovian conditioning and motivation at the three levels of computational, algorithmic and implementational analysis (Marr, 1982) that has emerged for instrumental conditioning and action (Dayan & Daw, 2008; Doll, Simon, & Daw, 2012). We take each of these levels of analysis in turn, reassembling and reshaping the pieces at the end.

2) The Computational Level

The computational level is concerned with the underlying nature of tasks, and the general logic or strategy involved in performing them (Marr, 1982). Here the task is prediction, and we consider model-based and model-free strategies.

A model-based strategy involves prospective cognition, formulating and pursuing explicit possible future scenarios based on internal representations of stimuli, situations and environmental circumstances[1] (Daw et al., 2005; de Wit & Dickinson, 2009; Sutton & Barto, 1998). This knowledge jointly constitutes a model, and supports the computation of value transformations when relevant conditions change (Tolman, 1948). Such models are straightforward to learn (i.e., acquisition is statistically efficient). However, making predictions can pose severe problems, since massive computations are required to perform the prospective cognition when this involves building and searching a tree of long-run possibilities extending far into the future. The leaves of the tree report predicted future outcomes whose values are also estimated by the model. Such estimates could be available in memory gained through corresponding past experience, e.g., actually tasting salt in novel state of sodium need. This process of acquiring new values through relevant state experiences is sometimes called UCS retasting (in the case of foods) or incentive learning in the more general instrumental case (Balleine & Dickinson, 1991). However, in cases such as the Dead Sea salt experiment that involve completely novel values and motivational states,

[1] In the reinforcement learning literature, what we are calling ‘circumstance’ is usually called a ‘state’. Here, we use the word circumstance to avoid confusion with motivational states (such as hunger and thirst). In general, motivational state forms part of the overall circumstance.

the tree-search estimates are inevitably constrained by what is not yet known (unless specific instructions or relevant generalization rules are prescribed in advance). That is, any experienced-derived search tree as yet contained no ‘leaves’ corresponding to a value of ‘nice saltiness’. Only nasty memories of intense saltiness were available. A new leaf would be required somehow to bud.

The other computational strategy is model-free. This is retrospective in the sense of operating purely using cached values accumulated incrementally through repeated experience (Daw et al., 2005; Dickinson & Balleine; Doya, 1999; Sutton & Barto, 1998), typically via a temporal difference prediction error (Sutton, 1988). Such model free processes must make their future estimates based on reward values that have been encountered in the past, rather than estimating the possible future. In the salt experiment above, therefore, the cached CS prediction error value would have been negative in the new appetite state, as it had been in past CS-UCS learning experiences. Model-free predictions are free of any content other than value, and are unaffected if the environment or the individual’s state suddenly changes since the past associations were learned -- at least until new learning coming from re-encounters with CS and UCS in the new state has adjusted the contents of the cache. Model-free algorithms such as temporal difference learning make predictions of the long-run values of circumstances, i.e., of the same quantities for which model-based learning builds a tree. They achieve this by bootstrapping – i.e., substituting current, possibly incorrect, estimates of the long-run worth for true values or samples thereof. Model-free estimation is statistically inefficient, because of this bootstrapping, since as at the outset of learning, the estimates used are themselves inaccurate. However model-free values are immediately available without the need for complex calculations.

The very different statistical and computational properties (Dayan & Daw, 2008) of model-based versus model-free strategies are a good reason to have both in the same brain. But when they co-exist, the two strategies can produce values that disagree (Dickinson & Balleine, 2002, 2010). Such discrepancies might be reconciled or resolved in various ways, for instance according to the relative uncertainties of the systems (Daw et al., 2005). So for example, a model-based strategy might dominate in early instrumental trials, when its superior statistical efficiency outweighs noise associated with the complex calculations, but a model-free strategy might dominate once learning is sufficient to have overcome the statistical inefficiency of bootstrapping. However, there are also intermediate points between the strategies that, at least in an instrumental context, are under active investigation, from viewpoints both theoretical (Dayan, 1993; Dezfouli & Balleine, 2012; Doll et al., 2012; Keramati, Dezfouli, & Piray, 2011; Pezzulo, Rigoli, & Chersi, 2013; Sutton & Barto, 1998) and empirical (Daw et al., 2011; Gershman, Markman, & Otto, 2012; Simon & Daw, 2011). In particular, model-based predictions might train model-free predictions offline (e.g., during quiet wakefulness or sleep; (Foster & Wilson, 2006, 2007) or on-line (Doll, Jacobs, Sanfey, & Frank, 2009; Gershman et al., 2012), or by providing prediction errors that can directly be used (Daw et al., 2011).

Distinguishing Pavlovian model-free from model-based

The computational literature has often assumed that Pavlovian learning is purely model-free (Montague, Dayan, & Sejnowski, 1996), similar to stimulus-response habits (Suri & Schultz, 1999). By contrast, we suggest here that a model-based computation is required to encompass the full range of evidence concerning Pavlovian learning and prediction. Our chief reason is that results from the Dead Sea salt experiment and others cited above hint at a crucial model-based feature: the computation must possess information about the sensory/perceptual identity of Pavlovian outcomes, distinct from mere previous values. Identity information is necessary to predict appropriately the value of a truly novel UCS that has never yet been experienced (e.g., an intense saltiness sensation as UCS identity, distinct from previous nastiness or niceness), and to apply that value change selectively to the appropriate CS (i.e., the salt-associated lever) without altering responses to other CSs (i.e., either the sucrose-associated lever or the control CS lever that had been associated with nothing). Identity information is also the most basic expression of a model–based mechanism that predicts an outcome rather than just carrying forward a previously cached value. However, as noted above, an identity prediction does not by itself suffice; it must also be connected to the modulation of value by the current physiological state, so that the saltiness representation of the UCS associated with CS could be predicted to have positive value in a way that would make CS become attractive and appropriately ‘wanted’. This predictive transformation is tricky since the taste outcome’s value had always been disgusting in the past. In particular, we must ask how this Pavlovian value computation is sensitive to the current brain-body state, even if novel, as the empirical results show it is.

We note that several straightforward ways of making a CS value computation sensitive to current state can be ruled out. For example, UCS re-tasting could have allowed the outcome to have been experienced as positively ‘liked’ rather than as disgusting, which would have updated any cognitive model-based representations derived from value experiences (Balleine & Dickinson, 1991; Dickinson & Balleine, 2010). But in the actual experiment, prior to the crucial CS test, neither the appetite nor the resulting pleasant value of the saltiness UCS had been experienced. Ensuring this novelty was one of the key intents of this experiment; it would be harder to guarantee with food satiety, for instance, since through alliesthesia, the subjects may have experience of eating food whilst relatively sated at the end of sustained bouts of feeding. In the Dead Sea experiment, from a computational view, the new value worth could only be inferred, i.e., recomputed a new based on internal representations of both the saltiness outcome and the novel motivational state relevant to future value. We will have to turn to alternative methods of computation that go beyond mere recall.

3) The Algorithmic Level

The algorithmic level concerns the procedures and representations that underpin computational strategies (Marr, 1982). Psychologically, this is clearest for instrumental conditioning (Daw et al., 2005; Dickinson & Balleine, 2002; Doya, 1999), with a rather detailed understanding of model-free temporal difference learning (Sutton, 1988), and a variety of suggestions for the nature of model-based calculations (Keramati et al., 2011; Pezzulo et al., 2013; Sutton & Barto, 1998).

There are two main algorithmic issues for Pavlovian conditioning: the first concerns the nature of the predictions themselves; the second, how those predictions are translated into behaviour. Our focus is on the former; however, we first touch on general aspects of the latter, since Pavlovian responses to CSs are how the predictions are assessed.

Pavlovian responses and incentive salience

CSs that predict appetitive and aversive outcomes elicit a range of conditioned responses. Appetitive predictors ordinarily become attributed with incentive salience during original reward learning, mediated neurobiologically by brain mesolimbic systems. The Pavlovian attribution of incentive salience let a targeted CS elicit surges of motivation that make that CS and its UCS temporarily more ‘wanted’ (Flagel et al., 2011; Mahler & Berridge, 2012; Robinson & Berridge, 2013; Robinson & Berridge, 1993). Incentive salience attributed to a CS can direct motivated behaviour toward that CS object or location, as in the Dead Sea salt experiment (with an intensity level that is dynamically modifiable by state-dependent changes in brain mesolimbic reactivity to Pavlovian stimuli (Flagel et al., 2011; Saunders & Robinson, 2012; Yager & Robinson, 2013)). Incentive salience attributed to the internal representation of a UCS associated with an external CS can also spur instrumental motivation to obtain that UCS as forms of what is known as Pavlovian to instrumental-transfer (PIT) (Colwill & Rescorla, 1988; Dickinson & Balleine, 2002; Dickinson & Dawson, 1987; Estes, 1943; Holland, 2004; Lovibond, 1981, 1983; Mahler & Berridge, 2012; Pecina & Berridge, 2013; Rescorla & Solomon, 1967). One form is specific PIT, when instrumental actions are directed at exactly the same appetitive outcome that the CS predicts (e.g., the same sugar pellets are Pavlovian UCSs and instrumental rewards). Another form is general PIT, in which instrumental effort is directed to a different outcome from the UCS associated with CS (though the outcome will generally be a related one, such as when a CS associated with one food spurs effort to obtain another food). In both cases, the CS spurs a burst of increased motivated effort, even though the CS may never previously have been associated with the instrumental behaviour (i.e., there exists no stimulus-response habit or association between CS and instrumental action).

An aversive CS which predicts an outcome such as a shock UCS, may elicit freezing; it may also suppress any ongoing, appetitively-directed, instrumental responding for food or another reward (Estes & Skinner, 1941; Killcross, Robbins, & Everitt, 1997). This can be seen as an aversive form of general PIT. Pavlovian anticipation of future punishments has further been suggested to lead to pruning of the model-based tree of future possibilities, potentially leading to suboptimal model-based evaluation (Dayan & Huys, 2008; Huys et al., 2012). Pruning is an influence over internal, cognitive, actions, rather than external, overt ones (Dayan, 2012).

Pavlovian values

As noted, the Dead Sea salt experiment suggests that the *identity* of the UCS (i.e., its saltiness) is predicted by the CS, distinct from the associated previous *values* (i.e., disgustingness). Once such a model-based mechanism is posited for Pavlovian learning it may be recognized as potentially playing a role in many CS-triggered incentive responses. Related UCS identity representations of sucrose reward, drug reward, etc. all might be implicated in Pavlovian CS amplifications of motivation induced by many neurobiological/physiological manipulations, ranging from permanent drug-induced sensitization to sudden brain stimulations of mesolimbic brain structures that magnify cuetriggered ‘wanting’ (e.g., dopamine/opioid drug stimulation of amygdala or nucleus accumbens).

In the Dead Sea salt experiment, the consequences of the prediction of identity can go one stage further, to conditioned alliesthesia (Toates, 1986), in which the CS is subject to the same physiological modulation as its UCS. Indeed, the Pavlovian lever/sound CS presentation sometimes elicited positive orofacial ‘liking’ reactions in the new appetite state, much as the taste of salt UCS itself later would on the same day (Robinson & Berridge, 2013). By contrast, if a model-free or pure valence-based Pavlovian mechanism had controlled responding, the mechanism would have continued to generate only disgust gapes and avoidance of the lever CS. Model-based control is also consistent with findings that Pavlovian blocking (i.e., the ability of an already-learned CS to prevent new learning to a second CS that begins to be simultaneously paired with the same UCS) dissipates when the identity of the blocked CS’s UCS changes, but its valence remains matched (McDannald, Lucantonio, Burke, Niv, & Schoenbaum, 2011).

However, CS revaluation is not ubiquitous. For example, sometimes in devaluation experiments, an originally appetitive CS persists in stimulating appetitive efforts after its UCS has been made worthless, consistent with model-free evaluation. This is especially well documented in taste aversion conditioning experiments involving overtraining of the food-seeking response prior to UCS devaluation. Although we discuss some differences later, there are further similarities between Pavlovian and instrumental effects of extensive training. For instance, it has been observed (Holland et al., 2008) that the predictive capacities of first-order appetitive CSs (which are directly associated with UCSs) are immediately affected by UCS revaluation, whereas second-order CSs (whose associations are established via first order CSs) are less influenced (Holland & Rescorla, 1975; Rescorla, 1973; Rescorla, 1974). Such results suggested that first order CSs establish stimulusstimulus associations (i.e., identity predictions), whereas second order CSs instead directly elicit responses engendered during conditioning (via stimulus-response associations). Indeed a related gradient of increasing CS resistance may apply to from UCS-proximal to distal CSs in Pavlovian serial associations, and from outcome-proximal to distal actions in instrumental circumstances (Balleine, Garner, Gonzalez, & Dickinson, 1995; Corbit & Balleine, 2003; Smith, Berridge, & Aldridge, 2011; Tindell, Berridge, Zhang, Peciña, & Aldridge, 2005).

The characteristic of model-free values of being tied to value but not identity of any specific outcome is especially evident in other paradigms. For instance in some situations, the absence of a negative-valenced event may be treated by an individual as similar to the occurrence of a positive-valenced event (Dickinson & Balleine, 2002; Holland, 2004) (Dickinson & Dearing, 1979; Ganesan & Pearce, 1988).

Pavlovian models

Having shown that model-based Pavlovian prediction can occur, we next consider what sort of model might be involved.

1. Stimulus substitution as the most elemental model-based mechanism—One simple form of model-based or sensory prediction that has not been considered from an instrumental viewpoint is Pavlov’s original notion of stimulus substitution. In this, the predicting CS comes to take on for the subject at least some of the sensory properties or qualities of the UCS it predicts, via direct activation of UCS-appropriate brain sensory regions (Pavlov, 1927); and the CS could then naturally come to take on some of the outcome’s incentive properties. Something akin to stimulus substitution is suggested when responses that are directed to the CSs resemble responses to the UCS (for instance pigeons pecking a key in a different way when it predicts food rather than water; (Jenkins & Moore, 1973)), and perhaps also some aspects of the progressive instinctive drift evident in Pavlovian misbehaviour, in which subjects come to manipulate the CS in some crucial ways as if it shares properties with the UCS (e.g., as when a pig roots a small CS object in a way that would normally be directed at a UCS piece of food) (Breland & Breland, 1961; Dayan, Niv, Seymour, & Daw, 2006). Note, though that substitution is never complete or literal: the CS is never actually mistaken for its UCS, and instead the nature of a Pavlovian response is always channelled by the CS identity as well as the UCS identity (Holland, 1977; Tomie, 1996). For example, a hungry rat that learns that the sudden appearance of another rat as CS predicts quick arrival of a food UCS, does not try to eat its fellow rat but rather responds with effusive approach, engagement, social grooming and related positive social behaviours (Timberlake & Grant, 1975). Such cases reflect CS substitution of UCS stimulus *incentive* properties rather than strict *identity* processes (Bindra, 1978; Toates, 1986). In short, the CS does not evoke a UCS hallucination.

Stimulus substitution might be seen as one of the simplest steps away from a pure valence expectation, involving a very simple associative (or mediating (Dwyer, Mackintosh, & Boakes, 1998) prediction; (Dickinson, 2012)). However, it is an efficient representational method to achieve some of the computational benefits of model-based predictions without requiring sophisticated previsioning machinery that depends on such processes as working memory. At the least it is an important hint that there may be more than one Pavlovian model-based mechanism.

2: Defocusing of modeled UCS identity representation—Another way to reconcile the facts that CSs can sometimes admit instantaneous revaluation (as in the Dead sea salt study), yet sometimes resist it (as in overtraining prior to taste aversion in the studies mentioned above), is to view the representation of the predicted UCS as flexible, and able to change over extended learning. We call this hypothesized process model-based UCS *defocusing*. For instance, over the course of extensive training, the UCS representation might become generalized, *blurred* or otherwise partially merged with representations of other related outcomes. This defocusing might lead to simple representations that afford generalization by dropping or de-emphasizing some of the particular sensory details of the UCS. Defocusing would be similar to basic concept learning of a category that contains multiple exemplars, such as of ‘tasty food’, whose representation evolves to be distinct from the unique identity of any particular example.

For a more intuitive view of defocusing, imagine the common experience of walking down a street as mealtime approaches and suddenly encountering the odor of food cooking inside a nearby building. Usually you guess the identity of what is being cooked, but sometimes you cannot. The food smell may be too complex or subtle or unfamiliar to recognize the precise identity of its UCS. In such a case, you have merely a defocused representation of the food UCS. But still you might feel suddenly as hungry as if you knew the UCS identity, and perhaps quite willing – even eager -- at that moment to eat whatever it is that you smell despite the lack of focus or any detailed identity knowledge in your UCS representation.

The implications of UCS defocusing are quite profound for the interpretation of experiments into devaluation insensitivity. Instead of resulting exclusively from model-free or habitual control as result of overtraining, persistence of responding to a CS could at least partly remain model-based, But if extensive training led a model’s representation of the UCS outcome to defocus, that defocused representation might escape any devaluation that depended on recalling the UCS’s precise sensory-identity details (e.g., Pavlovian taste aversion conditioning). The defocused representation could still support appetitive responding (at least until the UCS was actually obtained), despite the reduction in value of the actual UCS -- of which the subject might still show full awareness if tested differently. Thus dropping the particular identity taste representation of, say, fresh watermelon CS that has been paired with visceral illness as UCS, may leave a vaguer representation of juicy pleasant food that could still motivate appetitive effort until the previously delicious watermelon is finally retasted as now disgusting (Balleine & Dickinson, 1991; Dickinson & Balleine, 2010). This defocusing effect might especially result when manipulations used to revalue an outcome are essentially associative or learned, as distinguished from the physiological manipulation by appetite states, drugs, or brain activations that might more directly change CS value in parallel with UCS value, similar to the CS result for Dead sea saltiness (Berridge, 2012; Zhang et al. 2009). That difference may be because associative revaluations (e.g., Pavlovian taste aversions) layer on multiple and competing associations to the same food UCS, whereas physiological/brain states (e.g., salt appetite; addictive drugs) may more directly engage revaluation circuitry, and perhaps more readily revalue a CS’s ability to trigger incentive salience (Berridge, 2012).

The suggestion that defocusing occurs for predictions of a UCS should not be seen as contradicting our main proposition that the sensory identity of outcomes is key to understanding model-based Pavlovian learning and motivation. Instead, defocusing is associated with the development of a sophisticated, likely hierarchical, representation of the UCS and model-based predictions thereof, that admits an enriched set of multiple inferences and predictions, arranged along a spectrum of abstraction. For Pavlovian reward or threat exemplars, a variety of defocused or categorical UCS representations might exist: tasty foods, quenching drinks, sexual incentives, arousing drug reward states (e.g., amphetamine and cocaine), hedonic/calming drug reward states (e.g., heroin and morphine), painful events, and so on. These could be arranged in further hierarchical layers.

The details of how this spectrum is built need future clarification. However, it could proceed along the lines of unsupervised learning models for the development of cortical representations of general sensory input (Hinton & Ghahramani, 1997). Or it could be viewed as akin to the mechanisms of cognitive abstraction in declarative model-based systems, such as for a category of percepts (e.g., chairs in general) derived from several specific exemplars (e.g., particular chairs). Even pigeons can form perceptual abstractions, such as visual categories of pictures that contain images of trees or people, as relatively generalized concepts (Herrnstein, 1990).

Defocusing might also apply to Pavlovian representations of reward that influence instrumental behavior, such as in general PIT when presenting an appetitive CS spurs a burst of instrumental effort to obtain other rewards (but rewards that are usually categorically similar to the CS's UCS: e.g., tasty foods). Rather than depending on pure, model-free expectations of value, which is the conventional account of general PIT, this could depend on a model-based, but defocused, abstract, UCS prediction. For example, a CS for an ingestive UCS might trigger in different tests (a) specific PIT for its own UCS food, supported by a highly detailed representation of a reward's unique sensory identity (e.g., a saltiness representation for the Dead Sea salt CS transformation). The food CS might also trigger (b) a defocused, model-based, PIT for a different food UCS based on a more abstract representation similar to a basic concept (e.g., a tasty food lacking sensory details that produces persistent 'miswanting' after specific UCS devaluation). This defocused model would produce general PIT patterns of CS-triggered motivation for other food UCSs that belong to the same defocused class as its own UCS, but would not do so for categorically different UCSs that are quite different (e.g., non-food rewards such as noncaloric liquids, drugs, sex, etc.). Next, (c) there could be a nearly completely defocused representation of an outcome as simply having good or bad valence (allowing predicted omission of a good or bad outcome to be treated similarly to the predicted occurrence of a bad or good outcome respectively; as reflected in some tests of associative blocking). This would be close to (d) a true model-free general PIT for a non-ingestive reward, such as drug reward, sex reward, etc. Both (c) and (d) would generate equal intensities of general PIT for other food UCSs and for non-ingestive UCSs. However, (c) might still retain other model-based features that could be exposed by different tests. Indeed, future PIT experiments might usefully explore the possibility that there are multiple simultaneous representations for the same outcome, but at different degrees of defocusing. One way to do this would be to manipulate physiological states, for instance of hunger versus thirst, and then extend the range of instrumental choices in PIT experiments to include multiple UCSs belonging to different categories (e.g., food vs nonfood rewards), and modulating CS values via relevant vs irrelevant appetites. Such PIT experiments could make more evident the difference between model-free and defocused model-based predictions, and also elucidate the representational hierarchy for the latter.

One might wonder whether the most defocused or abstract UCS prediction could be just the same as pure, model-free, value. There are reasons to think not: the key distinction is that the range and form of generalization that underpins defocusing can be manipulated by information that is presented in contexts outside the precise learning problem. Take the case we mentioned above of smelling food whilst out walking. One could learn from a newspaper report that food sold on the street in London, say, is unhygienic. Such information might take the London UCS out of the generalization class of other street food, and perhaps reduce the motivating value of the CS scent of cooked food while walking London.

Extended training studies by Holland (Holland, 2004) assessing PIT after devaluation of the UCS (see also earlier examples of persistence after devaluation such as (Wilson, Sherman, & Holman, 1981)) might be reinterpreted as a concrete example of defocusing. As expected from the above, extending training rendered the instrumental response resistant to devaluation. More surprising, though, UCS devaluation also failed to reduce specific PIT, the boost to the vigor of instrumental actions aimed at obtaining the same identical UCS as predicted by the Pavlovian CS. That is, presenting the CS associated with a devalued UCS food still enhanced effort on the instrumental response which previously obtained that same food (the PIT test was conducted in extinction, without food outcomes actually being delivered), even though proximal conditioned responses to the CS, such as head entries into the food dish, were reduced by the devaluation. This would be consistent with multiple simultaneous representations of the UCS, with the Pavlovian one that guided instrumental behaviour being defocused when accessed by instrumental learning systems, and so unaffected by the particular, identity-specific, devaluation procedure.

Defocusing or loss of UCS identity might also relate to Tolman’s (Tolman, 1949; Tolman, 1955) interpretation of the original demonstrations that extended overtraining could induce resistance to subsequent UCS devaluation (sometimes called ‘habits’ for that reason). Those demonstrations showed that a suddenly hungry rat, which had always previously been trained while thirsty, continued to seek the location of a water reward in a maze, and continued to ignore the location of an alternative food reward that now ought to be valuable (Thistlethwaite, 1952). Tolman thought that this might involve a “narrowing” of the cognitive map. In his own words, “even though a rat’s drive be changed from thirst to hunger, his need-push may not, for a while, change correspondingly. The water place may still be valenced, even though the drive as measured by our original definition is now one of hunger. In other words, under some conditions, rats, and men too, do not seem to have need-pushes which correspond to their actual drives (and also I would remark, parenthetically, they may often experiences valences that do not correspond to their actual values).” (p. 368, (Tolman, 1949)). Although habit-theorists might be tempted to view the lagging ‘need-push’ as a model-free propensity, an alternative based on defocusing is to view it as a defocused persistence of the cognitive representation of the value of act-outcome value in the new state, until reinstructed by value experiences relevant to that state (e.g., food becoming more valuable during hunger), all contained in a model-based or cognitive-style representation (Daw et al., 2005; Dickinson & Balleine, 2002; Doya, 1999). Such re-tasting opportunities lead the rat to subsequently switch to seeking food whenever in the hunger state and to not persist in seeking water in the maze (Thistlethwaite, 1952). Tolman himself provided a rather model-based account of what he meant in terms of expectancies and cognitive maps in a related article: namely that thirsty overtraining with the water reward “interfered with activations of the appropriate scannings and consequent additional discriminations and expectancies necessary for the development of a pragramatic performance vector with respect to the food. The already aroused strong approach-towater performance vector tended, as I have put it elsewhere, to narrow the rat’s ‘cognitive maps’.” (p.36; (Tolman, 1955)). Although not identical to UCS defocusing, a narrowing of a cognitive map that prevents appropriate scanning of reward expectancies to assess new value might best be viewed in model-based terms.

Such concepts of narrowing the cognitive map or defocusing make it harder to distinguish between model-free and model-based control, since they argue that the model-based system can suffer from a form of pathology that makes its predictions resemble those of a model-free system. However, the concepts do not challenge the fundamental distinction between the two systems; rather they invite a more discriminative set of experiments, perhaps of the flavour of those described above.

How to characterize Pavlovian model-based evaluation computationally?—A major computational challenge regarding Pavlovian valuation is to capture the change in CS value in algorithmic form. This challenge has yet to be fully met. In fact, a primary purpose of our writing this paper is to inspire further attempts to develop better computational models for Pavlovian CS-triggered motivations in future. As an initial step, Zhang and colleagues (Zhang, Berridge, Tindell, Smith, & Aldridge, 2009) proposed a phenomenological model of CS-triggered incentive salience, as the motivational transform of CS value from a previously learned cache of prediction errors (described in the Appendix). But, as those authors themselves agreed, much more remains to be done.

According to the Zhang account, the cached value of previous UCS encounters is changed by a physiological-neurobiological factor called kappa that reflects the current brain/body state of the individual (whether the state is novel or familiar). The current kappa value multiplies or logarithmically transforms a temporal difference cache associated with a CS when the cue is re-encountered. That transformation operation would be mediated by the mesocorticolimbic activations that produce incentive salience. The Zhang model succeeds in describing quantitatively the value transformations induced by salt appetite, other appetites and satieties, drug-induced priming of motivation, etc. addiction-related sensitization. However, the Zhang description is purely external to the mechanism in the sense that a kappa modification of a UCS value memory associated with CS captures the transformed motivation output, but does not provide any hypothesis about the internal algorithmic process by which the transformation is achieved. Essentially, the Zhang model shows how violence must be done to any pre-existing model-free cache of learned values, such as that accumulated by a temporal difference mechanism, in order to achieve the newly transformed value that can appear in a new state. However, our whole point here is that the accomplishment of such Pavlovian state transformations essentially requires a model-based mechanism, not a model-free one, implying that a quite different computational approach will eventually be required. A comprehensive algorithmic version of the model-based Pavlovian computation has yet to be proposed. We hope better candidates will be proposed in coming years to help fill this important gap.

4) The Implementational Level

Marr's implementational level concerns the way that the algorithms and representations are physically realized in the brain (Marr, 1982). There is a wealth of data from rodents and human and non-human primates as to neural systems involved in model-based and model-free instrumental systems; given the observation that Pavlovian systems require and exploit predictions of long-run utility in closely related ways, one might hope that these results would generalize.

Very crudely, brain regions such as prefrontal cortex and the dorsomedial striatum, as well as the hippocampus and the default network might be involved in model-based prediction and control (Hassabis, Kumaran, Vann, & Maguire, 2007; Johnson & Redish, 2007; Pfeiffer & Foster, 2013; Schacter, Addis, & Buckner, 2008; Schacter et al., 2012; Spreng, Mar, & Kim, 2009; van der Meer, Johnson, Schmitzer-Torbert, & Redish, 2010). The dopamine system that originates in ventral tegmentum (VTA) and substantia nigra pars compacta (SNc), and its striatal targets, perhaps especially in dorsolateral neostriatum, have sometimes been suggested as being chiefly involved in model-free learning (Balleine, 2005; Daw et al., 2011; Dickinson & Balleine, 2002; Gläscher et al., 2010; Hikosaka et al., 1999; Killcross & Coutureau, 2003; Samejima, Ueda, Doya, & Kimura, 2005; Simon & Daw, 2011; Wunderlich et al., 2012). This has also been contested and will be examined below.

Some paradigms, notably Pavlovian-instrumental transfer (PIT), provide an additional and more selective view. It is known from rodents that there is a particular involvement of circuits linking the amygdala and the accumbens in PIT, with special roles for the basolateral nucleus of the amgydala and possibly the shell of the accumbens in specific PIT, which is the form of PIT related to model-based evaluation, and the central nucleus of the amygdala and possibly the core of the accumbens in general PIT, which some regard as closer to model-free evaluation (Balleine, 2005; Corbit & Balleine, 2005; Corbit, Janak, & Balleine, 2007; Hall, Parkinson, Connor, Dickinson, & Everitt, 2001; Holland & Gallagher, 2003; Mahler & Berridge, 2012). A related circuit has been implicated in human PIT (Bray, Rangel, Shimojo, Balleine, & O'Doherty, 2008; Prevost, Liljeholm, Tyszka, & O'Doherty, 2012; Talmi, Seymour, Dayan, & Dolan, 2008). Note, though our discussion above implying that general PIT might be reinterpreted as a form of defocused, model-based, specific PIT. Certainly general PIT undergoes similar transformations that enhance or suppress the ability of valued CSs to trigger 'wanting' surges in response to neurochemical stimulations of either nucleus accumbens (shell or core) or central amygdala (Dickinson, Smith, & Mirenowicz, 2000; Mahler & Berridge, 2012; Pecina & Berridge, 2013; Wassum, Ostlund, Balleine, & Maidment, 2011; Wyvell & Berridge, 2000). Defocusing would force us to draw rather different conclusions from these various anatomical studies into the substrates of different control systems (Balleine, 2005).

Where in the brain does Pavlovian motivational revaluation of CS occur?

The answer must accommodate the ability of Pavlovian model-based systems to calculate the values of predicted outcomes under current motivational states. Some might suggest that such prospective revaluation should occur at a cortical level, perhaps involving ventromedial regions of prefrontal cortex. There is quite a wealth of evidence that orbitofrontal and related prefrontal areas are involved in model-based predictions of the values associated with stimuli and their predictors, in cases when value has been obtained by previous experiences in relevant states (Boorman, Behrens, Woolrich, & Rushworth, 2009; Camille, Tsuchida, & Fellows, 2011; Jones et al., 2012; McDannald et al., 2012; O'Doherty, 2011) or even in the apparently purely model-based task of assigning value to imagined foods (Barron, Dolan & Behrens,2013). These areas are apparently not so involved instrumentally in directly assigning pre-experienced values to current actions (Camille et al., 2011; O'Doherty, 2011), although they can potentially support stimulus-based rather than actionbased choice (O'Doherty, 2011).

The orbitofrontal cortex was one of the areas in the rat that was found in the Dead sea salt experiment to have greatly up-regulated activity in test trials following induction of salt appetite, when the CS lever was re-encountered as being attractive in the new appetite state (Robinson & Berridge, 2013). However, the fact that animals whose neocortex has been surgically removed can still show revaluation of learned relations in the face of a new salt appetite (Wirsig & Grill, 1982) suggests that cortex (including orbitofrontal cortex) is at least not necessary for this revaluation. Pre-programmed sub-cortical sophistication for Pavlovian revaluation could be highly adaptive in realizing the most fundamental needs of an organism; the questions then become the range of outcomes for which subcortical transformation is possible (e.g., different sorts of natural appetite/thirst states, drug-induced states or beyond), and the identity of the regulatory circuitry that interfaces with mesolimbic circuitry at least for the monitoring of natural need states (hypothalamus, etc.)(Berthoud & Morrison, 2008; Gao & Horvath, 2008; Krause & Sakai, 2007). The potentially subcortical nature of Pavlovian motivation transformations may have implications for the degree of sophistication in the model-based previsioning -- is it purely temporally local for immediately-paired CS-UCS associations (e.g., depending on forms of stimulus substitution), or can it also bridge stimuli and time, as in richer forms of secondary conditioning? How defocused are these sub-cortical predictions? Can they contribute at all to instrumental model-based evaluation?

Role of mesolimbic dopamine?

Perhaps the most contentious contributor to evaluation is the neuromodulator dopamine. Dopamine neurons in the midbrain VTA project to the nucleus accumbens (ventral striatum) and the prefrontal cortex, and dopamine neurons in the adjacent SNc project to the neostriatum (dorsal striatum), with various subregional patterns of further localization. Many of these dopamine systems have been implicated in reward, though argument continues over precisely which reward-related functions are performed. As a brief summary of evidence, dopamine neurons projecting to nucleus accumbens and neostriatum respond similarly to rewards and to learned Pavlovian and instrumental cues (Montague et al., 1996; Morris, Nevet, Arkadir, Vaadia, & Bergman, 2006; Roesch, Calu, & Schoenbaum, 2007; Schultz, 1998, 2006; Schultz, Dayan, & Montague, 1997), and dopamine release in animals and humans is linked to rewards and cues in both striatum and nucleus accumbens (Boileau et al., 2006; Darvas & Palmiter, 2010; de la Fuente-Fernandez et al., 2002; Kishida et al., 2011; Phillips, Stuber, Heien, Wightman, & Carelli, 2003; Roitman, Stuber, Phillips, Wightman, & Carelli, 2004; Volkow, Wang, Fowler, & Tomasi, 2012; Wanat, Willuhn, Clark, & Phillips, 2009; Wise, 2009; Zaghloul et al., 2009). Animals readily learn to emit actions in order to activate dopamine neurons in the VTA and SNc (Nieh, Kim, Namburi, & Tye, 2013; Rossi, Sukharnikova, Hayrapetyan, Yang, & Yin, 2013; Witten et al., 2011). There is also a rich pattern of connections from these structures to the ventral pallidum and to them from the amygdala, as well as with other subcortical nuclei such as the lateral habenula and rostromedial tegmental nucleus (RMTg), the serotonergic raphe nucleus; and also pathways linking them to the hypothalamus (Moore & Bloom, 1978; Swanson, 1982). Further, the activity of dopaminergic cells, their release of dopamine, and/or the longevity of the neuromodulator at its targets are modulated by almost all addictive drugs (Hyman, Malenka, & Nestler, 2006; Koob & Volkow, 2010; Volkow et al., 2012). Finally, repeated exposure to addictive drugs can more permanently sensitize dopamine-related circuits in susceptible individuals in ways that enhance neural responses to learned reward cues (Leyton & Vezina, 2012; Robinson & Berridge, 2008; Robinson & Kolb, 2004; Thomas, Kalivas, & Shaham, 2008; Vezina & Leyton, 2009; Wolf & Ferrario, 2010).

Dopaminergic neurons and many of their targets are modulated by neuropeptide and hormone signals such as corticotropin releasing factor or ghrelin released by the hypothalamus or the periphery that can report on current states of stress or appetite (e.g., feeding-related) motivational state (Korotkova, Brown, Sergeeva, Ponomarenko, & Haas, 2006; Zigman, Jones, Lee, Saper, & Elmquist, 2006). The VTA and nucleus accumbens were notable among the structures recruited at the moment of salt cue re-encounter during appetite in the Robinson and Berridge study, raising the possibility of dopamine activations as part of the mechanism for sudden CS transformation from repulsive to ‘wanted’. That possibility is made plausible because elevation of dopamine levels in nucleus accumbens shell or core directly enhances the degree of ‘wanting’ triggered by reward CS above any previously learned levels in general PIT behaviour, and in limbic neuronal firing to the CS in ventral pallidum, a chief output structure for nucleus accumbens (Pecina & Berridge, 2013; Smith et al., 2011; Tindell et al., 2005).

Much computational effort in the past decade focused on understanding these roles of dopamine has focused on its possible involvement in model-free learning, especially in the form of a temporal difference prediction error for future reward which the phasic activity of dopamine neurons strikingly resembles (Barto, 1995; Berridge, 2007; Mahler & Berridge, 2012; Montague et al., 1996; Schultz, 2006; Schultz et al., 1997). One view suggests that a phasic dopamine pulse is the key teaching signal for model-free prediction and action learning, as in one of reinforcement learning’s model-free learning methods: the actor critic; (Barto, Sutton, & Anderson, 1983), Q-learning (Roesch et al., 2007; Watkins, 1989), or SARSA (Morris et al., 2006; Rummery & Niranjan, 1994). The same dopamine signal can act to realise incentive salience, either as cached value or as Dead sea type transformation (McClure, Daw, & Montague, 2003; Zhang et al., 2009).The actor-critic is of particular interest (Li & Daw, 2011), since, as studied in conditioned reinforcement (Mackintosh, 1983) or escape from fear (McAllister, McAllister, Hampton, & Scoles, 1980), it separates out circumstance-based predictions (in the critic) from action-contingency (in the actor). There is evidence for circumstance-based and action-based prediction errors in distinct parts of the striatum (O’Doherty et al, 2004), although the action-based errors were value based (associated with a variant of a state-action prediction called a Q-value), rather than purely action based (as in the actor portion of the actor critic (Li & Daw, 2011)). The fact that the critic evaluates circumstances rather than actions under particular circumstances makes it a natural candidate as a model-free predictor that can support both Pavlovian and instrumental conditioning, although it remains to be seen if there are also dual value- and action-based routes to Pavlovian actions, with the latter being a stamped-in stimulus response mapping (analogous to the instrumental actor). Some complications do arise from the spatial heterogeneity for valence coding that has particularly been observed in the VTA (but see (Matsumoto & Hikosaka, 2009)), with one group of dopamine neurons being excited by unexpected punishments (rather than being suppressed, as might naively be expected for a prediction error for reward) (Brischoux, Chakraborty, Brierley, & Ungless, 2009; Lammel, Lim, & Malenka, 2013; Lammel et al., 2012).

Equally, some tonic aspects of dopamine release have been suggested to mediate the vigour of action (for instance, reducing reaction times) or the exertion of effort (Niv, Daw, Joel, & Dayan, 2007; Salamone & Correa, 2002). Normal levels of tonic dopamine release are necessary for realizing both appetitive and aversive preparatory motivations elicited by stimulation of the accumbens (Faure, Reynolds, Richard, & Berridge, 2008; Richard & Berridge, 2011), and are involved in at least amplifying phasic bursts of motivation triggered by CS encounters, such as in appetitive PIT (Corbit et al., 2007; Murschall & Hauber, 2006). Appealingly for model-free learning theorists, the straightforward integration over time of the phasic prediction error formally signals the average reward (Daw, Kakade, & Dayan, 2002) ; although tonic and phasic dopamine activity may be under somewhat separate control (Floresco, West, Ash, Moore, & Grace, 2003; Goto & Grace, 2005), and it is not clear whether there is also a model-based contribution to this average. The average reward reports the opportunity cost for the passage of time, and has thus been interpreted as key to the instrumental choice of vigor (Niv et al., 2007).

However, a model-free learning interpretation of dopamine mesolimbic function cannot be the whole story here either (Berridge, 2007, 2012). This is indicated by motivational transformation results such as those of the Dead Sea salt experiment and some others mentioned above. Thus, the highly significant upregulation in activity in VTA and target structures such as nucleus accumbens triggered by the CS following the induction of salt appetite suggest that dopamine release might also be dramatically greater, potentially licensing the excess behaviour directed towards the conditioned stimulus. We argued above that this revaluation is the preserve of Pavlovian model-based reasoning, and cannot be accomplished by a model-free system. This then suggests that dopamine release can actually reflect model-based evaluations rather than (at least only) model-free predictions. Similar conclusions might be drawn from the finding that inactivating the VTA disrupts both specific and general PIT (Corbit et al., 2007).

Does the involvement of dopamine in model-based CS evaluations, which had traditionally been thought of instrumentally as being associated with model-free calculations, again imply a critical difference between mechanisms of Pavlovian and instrumental model-based evaluation? There are reasons for thinking so. For instance, Dickinson and Balleine postulated that retasting the new value of an outcome in any novel motivational state is necessary for instrumental revaluation to discover its changed value (Dickinson & Balleine, 2010), and retasting of food while hungry in the maze was also able to revalue seeking of food in the original Thistlewaite (Thistlethwaite, 1952) water/food revaluation experiments discussed by Tolman (Tolman, 1949; Tolman, 1955). This form of instrumental incentive learning appears independent of dopamine, proceeding normally under dopamine receptor blockade (Dickinson & Balleine, 2010; Dickinson et al., 2000; Wassum et al., 2011). By contrast, in the Pavlovian case, revaluation does not require retasting, and is powerfully modulated by dopamine (the CS value is suppressed by blockade, and magnified in value by dopamine-stimulating drugs).

Further, it is endogenous features of an individual's dopamine systems that may be associated with the differences among individuals in the way they assign motivational value to a particular Pavlovian CS, such as a discrete distal cue that is highly predictive of reward UCS (Flagel et al., 2011; Saunders & Robinson, 2012; Yager & Robinson, 2013). For example, (Flagel et al., 2011) measured the release and role of dopamine in two groups of rats: 'sign-trackers', whose motivated responses directionally targeted to the discrete are Pavlovian CS, and 'goal-trackers', who appear to eschew targeting Pavlovian incentive salience to that CS, and instead approach only the dish that delivers UCS (potentially mediated also by instrumental expectations or by habits). Only the sign-trackers showed a substantial elevation in dopamine release to the predictive CS that attracted them; and their behaviour was most sensitively influenced by dopamine antagonists. If the goal-trackers are indeed more subject to instrumental model-based or to habitual model-free, control, then this absence of dopamine effects contrasts with the dopamine dependence of Pavlovian model-based control of incentive salience that we documented above.

In the end, despite these differences, the case is still open. Take, for instance, instrumental incentive learning. Daw and colleagues (Daw et al., 2005) suggested that the apparent requirement for the outcome to be retasted in order to see an effect of devaluation, could instead reflect involvement of model-free habits that compete with a more readily revalued model for behavioural control. Their idea is that instrumental model-based prospective evaluation, just like Pavlovian prospective evaluation, has access to the new value of the UCS. However, because of the change in motivational state, the instrumental evaluation is also aware that it is less certain about this value because novelty promotes uncertainty. Retasting duly reduces that uncertainty. By contrast, model-free predictions know about neither the new value nor the associated new uncertainty. Thus, if model-free and model-based systems compete according to their relative certainties, the model-free habit, and thus devaluation insensitivity, will dominate until re-tasting. However, Pavlovian model-based predictions might be less uncertain than instrumental ones, since they do not have to incorporate an assessment of the contingency between action and outcome, and so may more easily best their model-free counterparts. Thus, there might still only be one model-based knowledge system, but two control systems or different ways of translating knowledge into action: instrumental act-outcome performance (disrupted by uncertainty) and Pavlovian motivation (less affected by uncertainty in this instance). The route by which this model-based information takes its Pavlovian effects could involve corticolimbic inputs from prefrontal cortex to the midbrain dopamine system. Potential experiments might test this idea, for instance by manipulating such inputs and examining whether instant CS revaluation effects such as the Dead sea salt study still obtain.

More generally, there is increasing evidence for rich interactions between model-free and model-based predictions (Daw et al., 2011; Gershman et al., 2012; Simon & Daw, 2011). Then, for instance, the activation of VTA dopamine neurons to a saltiness-related CS, if observed in the salt appetite conditions, could arise from a model-based system as part of the way that this putatively trains model-free predictions (Doll et al., 2009; Foster & Wilson, 2006, 2007; Gershman et al., 2012). There may still be differences in detail -- for instance, with model-based influences over dopamine involving the VTA more than the SNc. These remain to be explored.

5) Synthesis

We started with an experimental example of instant CS revaluation in the light of prevailing motivational states that poses an important challenge to standard computational accounts of learning and performance in Pavlovian conditioning. Our suggested answer is in one way rather simple – directly importing model-based features that are standard explanations in instrumental conditioning into what have been sometimes treated as purely model-free Pavlovian systems, i.e., including for Pavlovian predictions what has long been recognized for instrumental predictions. The revaluation is exactly what a model-based system ideally could produce, i.e., reporting the current value of a predicted outcome. Key questions remaining include the circumstances under which this re-computation would seize control, the neural mechanisms responsible and how direct CS modulation is achieved without necessarily requiring tasting of an altered UCS.

However, looked at more closely, things get more interestingly complicated in at least two ways. First, the nature of the computations and algorithms underlying Pavlovian model-based predictions remain open for investigation and future modelling. We discussed evidence hinting that these might not be completely shared with instrumental model-based predictions. The apparently embellished scope of Pavlovian model-based calculation includes such things as instant revaluation, in both normal and decorticate subjects, putatively involving sensory-identity representations of UCS, and the possibility of defocusing that representation into a categorical one along the spectrum between specific and general predictions. These ideas enrich our picture of model-based systems (potentially even applying in some respects to instrumental model-based mechanisms).

Second, consider the conclusion that the results of the salt appetite experiment or other mesolimbic manipulations of cue-triggered incentive salience indeed depend on model-based calculations. This implies that Pavlovian model- and identity-based predictions burrow directly into what has previously been thought of as the neurobiological heart of model-free and purely valence-based predictions (i.e. a temporal difference prediction error mechanism), namely dopamine activity and release in nucleus accumbens and related mesostriatal and mesolimbic circuitry. It therefore becomes pressing to re-examine more closely the role of dopamine brain systems in reward learning and motivation. That might include tipping the balance between model-based and model-free Pavlovian predictions. Such issues might be studied, for instance using manipulations such as reversible pre- and infra-limbic lesions or dorsomedial and dorsolateral neostriatal manipulations (Balleine & O'Doherty, 2010; Difeliceantonio, Mabrouk, Kennedy, & Berridge, 2012; Killcross & Coutureau, 2003; Smith, Virkud, Deisseroth, & Graybiel, 2012) that have been so revealing for instrumental conditioning.

In summary, the current computational analysis invites a blurring between model-free and model-based systems and between Pavlovian and instrumental predictions. What is clearly left is that there is significant advantage to having structurally different methods of making predictions in a single brain; that there is a critical role for pre-programming in at least some methods for making predictions; that the attention to Pavlovian model-based predictions makes even more acute the question of the multifarious nature of exactly what might be predicted; and finally that all these issues are played out over a range of cortical and sub-cortical neural systems whose nature and rich interactions are presently becoming ever more apparent.

Acknowledgments

This manuscript began in a series of discussions between PD and KB while KB was a sabbatical visitor for two months at the Gatsby Computational Neuroscience Unit. We thank Ritwik Niyogi for the suggestion during that visit that the authors should write together about model-based Pavlovian control. We are also particularly grateful to him, Wayne Aldridge, Josh Berke, Todd Braver, Tony Dickinson, Peter Holland, Zeb Kurth-Nelson, Terry Robinson, Martin Sarter and Natalie Tronson for helpful comments on earlier drafts. Manuscript preparation and research from the authors' laboratories was supported by grants from the Gatsby Charitable Foundation to PD, and the NIH (DA015188 and MH63649) and University of Michigan to KB.

Bibliography

Anson JE, Bender L, Melvin KB. Sources of reinforcement in the establishment of self-punitive behavior. Journal of Comparative and Physiological Psychology. 1969; 67(3):376–380. [PubMed: 5787389]

Balleine BW. Asymmetrical interactions between thirst and hunger in Pavlovian-instrumental transfer. Quarterly Journal of Experimental Psychology B, Comparative & Physiological Psychology. 1994; 47(2):211–231.

Balleine BW. Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits. Physiology and Behavior. 2005; 86(5):717–730. [PubMed: 16257019]

Balleine BW, Dickinson A. Instrumental performance following reinforcer devaluation depends upon incentive learning. The Quarterly Journal of Experimental Psychology. 1991; 43(3):279–296.

Balleine BW, Garner C, Gonzalez F, Dickinson A. Motivational control of heterogeneous instrumental chains. Journal of Experimental Psychology: Animal Behavior Processes. 1995; 21(3):203.

Balleine BW, O'Doherty JP. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology. 2010; 35(1):48–69. [PubMed: 19776734]

Barto, AG. Adaptive critics and the basal ganglia. In: Houk, J.; Davis, J.; Beiser, D., editors. Models of Information Processing in the Basal Ganglia. Cambridge, MA: MIT Press; 1995. p. 215-232.

Barto AG, Sutton RS, Anderson CW. Neuronlike elements that can solve difficult learning control problems. IEEE Transactions on Systems, Man, and Cybernetics. 1983; 13(5):834–846.

Berridge KC. The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology (Berl). 2007; 191(3):391–431. [PubMed: 17072591]

Berridge KC, Schulkin J. Palatability shift of a salt-associated incentive during sodium depletion. Quarterly Journal of Experimental Psychology [b]. 1989; 41(2):121–138.

Berthoud HR, Morrison C. The brain, appetite, and obesity. Annu Rev Psychol. 2008; 59:55–92. [PubMed: 18154499]

Bindra D. How adaptive behavior is produced: a perceptual-motivation alternative to response reinforcement. Behavioral and Brain Sciences. 1978; 1:41–91.

Boileau I, Dagher A, Leyton M, Gunn RN, Baker GB, Diksic M, et al. Modeling sensitization to stimulants in humans: an [11C]raclopride/positron emission tomography study in healthy men. Arch Gen Psychiatry. 2006; 63(12):1386–1395. [PubMed: 17146013]

Boorman ED, Behrens TE, Woolrich MW, Rushworth MF. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. Neuron. 2009; 62(5): 733–743. [PubMed: 19524531]

Bouton ME, Moody EW. Memory processes in classical conditioning. Neuroscience and biobehavioral reviews. 2004; 28(7):663–674. [PubMed: 15555676]

Bray S, Rangel A, Shimojo S, Balleine BW, O’Doherty JP. The neural mechanisms underlying the influence of pavlovian cues on human decision making. Journal of Neuroscience. 2008; 28(22): 5861–5866. [PubMed: 18509047]

Breland K, Breland M. The misbehavior of organisms. American Psychologist. 1961; 16(9):681–684.

Brischoux F, Chakraborty S, Brierley DI, Ungless MA. Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(12):4894–4899. [PubMed: 19261850]

Bushong B, King LM, Camerer CF, Rangel A. Pavlovian processes in consumer choice: The physical presence of a good increases willingness-to-pay. The American Economic Review. 2010; 100(4): 1556–1571.

Camille N, Tsuchida A, Fellows LK. Double dissociation of stimulus-value and action-value learning in humans with orbitofrontal or anterior cingulate cortex damage. J Neurosci. 2011; 31(42):15048–15052. [PubMed: 22016538]

Campbell M, Hoane AJ Jr, Hsu F-h. Deep Blue. Artificial Intelligence. 2002; 134(1–2):57–83.

Colwill RM, Rescorla RA. Associations between the discriminative stimulus and the reinforcer in instrumental learning. Journal of experimental psychology Animal behavior processes. 1988; 14(2):155–164.

Corbit LH, Balleine BW. Instrumental and Pavlovian incentive processes have dissociable effects on components of a heterogeneous instrumental chain. J Exp Psychol Anim Behav Process. 2003; 29(2):99–106. [PubMed: 12735274]

Corbit LH, Balleine BW. Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of pavlovian-instrumental transfer. The Journal of neuroscience. 2005; 25(4):962–970. [PubMed: 15673677]

Corbit LH, Janak PH, Balleine BW. General and outcome-specific forms of Pavlovian-instrumental transfer: the effect of shifts in motivational state and inactivation of the ventral tegmental area. European Journal of Neuroscience. 2007; 26(11):3141–3149. [PubMed: 18005062]

Darvas M, Palmiter RD. Restricting dopaminergic signaling to either dorsolateral or medial striatum facilitates cognition. J Neurosci. 2010; 30(3):1158–1165. [PubMed: 20089924]

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans’ choices and striatal prediction errors. Neuron. 2011; 69(6):1204–1215. [PubMed: 21435563]

Daw ND, Kakade S, Dayan P. Opponent interactions between serotonin and dopamine. Neural Netw. 2002; 15(4–6):603–616. [PubMed: 12371515]

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

Dayan P. Improving generalization for temporal difference learning: The successor representation. Neural Computation. 1993; 5(4):613–624.

Dayan P, Daw ND. Decision theory, reinforcement learning, and the brain. Cognitive, Affective, & Behavioral Neuroscience. 2008; 8(4):429–453.

Dayan P, Huys QJM. Serotonin, inhibition, and negative mood. PLoS Computational Biology. 2008; 4(2):e4–e4. [PubMed: 18248087]

Dayan P, Niv Y, Seymour B, Daw ND. The misbehavior of value and the discipline of the will. Neural Netw. 2006; 19(8):1153–1160. [PubMed: 16938432]

de la Fuente-Fernandez R, Phillips AG, Zamburlini M, Sossi V, Calne DB, Ruth TJ, et al. Dopamine release in human ventral striatum and expectation of reward. Behavioural Brain Research. 2002; 136(2):359–363. [PubMed: 12429397]

de Wit S, Dickinson A. Associative theories of goal-directed behaviour: a case for animal-human translational models. Psychol Res. 2009; 73(4):463–476. [PubMed: 19350272]

Dezfouli A, Balleine BW. Habits, action sequences and reinforcement learning. Eur J Neurosci. 2012; 35(7):1036–1051. [PubMed: 22487034]

Dezfouli A, Balleine BW. Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. PLoS Comput Biol. 2013; 9(12):e1003364. [PubMed: 24339762]

Dickinson A. Re-examination of the role of the instrumental contingency in the sodium-appetite irrelevant incentive effect. Q J Exp Psychol B. 1986; 38(2):161–172. [PubMed: 3737979]

Dickinson A. Associative learning and animal cognition. Philos Trans R Soc Lond B Biol Sci. 2012; 367(1603):2733–2742. [PubMed: 22927572]

Dickinson, A.; Balleine, BW. The role of learning in motivation. In: Gallistel, CR., editor. Stevens' handbook of experimental psychology. Vol. 3. Wiley; 2002. p. 497-533.

Dickinson, A.; Balleine, BW. Hedonics: The Cognitive-Motivational Interface. In: Kringelbach, ML.; Berridge, KC., editors. Pleasures of the brain. Oxford, U.K: Oxford University Press; 2010. p. 74-84.

Dickinson A, Dawson GR. Pavlovian processes in the motivational control of instrumental performance. The Quarterly Journal of Experimental Psychology. 1987; 39(3):201–213.

Dickinson, A.; Dearing, MF. Appetitive-aversive interactions and inhibitory processes. In: Dickinson, A.; Boakes, RA., editors. Mech. Hillsdale, New Jersey: Lawrence Erlbaum; 1979. p. 203-231.Vol. Mechanisms of learning and motivation: A memorial volume to Jerzy Konorski

Dickinson A, Smith J, Mirenowicz J. Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. Behavioral Neuroscience. 2000; 114(3):468–468. [PubMed: 10883798]

Difeliceantonio AG, Mabrouk OS, Kennedy RT, Berridge KC. Enkephalin Surges in Dorsal Neostriatum as a Signal to Eat. Current Biology. 2012; 22(20):1918–1924. [PubMed: 23000149]

Doll BB, Jacobs WJ, Sanfey AG, Frank MJ. Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. Brain Research. 2009; 1299:74–94. [PubMed: 19595993]

Doll BB, Simon DA, Daw ND. The ubiquity of model-based reinforcement learning. Curr Opin Neurobiol. 2012; 22(6):1075–1081. [PubMed: 22959354]

Doya K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Netw. 1999; 12(7–8):961–974. [PubMed: 12662639]

Dwyer DM, Mackintosh NJ, Boakes RA. Simultaneous activation of the representations of absent cues results in the formation of an excitatory association between them. Journal of experimental psychology Animal behavior processes. 1998; 24(2):163–171.

Estes WK. DISCRIMINATIVE CONDITIONING. I. A DISCRIMINATIVE PROPERTY OF CONDITIONED ANTICIPATION. Journal of Experimental Psychology. 1943; 32:150–155.

Estes WK, Skinner BF. Some quantitative properties of anxiety. Journal of Experimental Psychology. 1941; 29(5):390–400.

Faure A, Reynolds SM, Richard JM, Berridge KC. Mesolimbic dopamine in desire and dread: enabling motivation to be generated by localized glutamate disruptions in nucleus accumbens. Journal of Neuroscience. 2008; 28(28):7184–7192. [PubMed: 18614688]

Fermin A, Yoshida T, Ito M, Yoshimoto J, Doya K. Evidence for model-based action planning in a sequential finger movement task. J Mot Behav. 2010; 42(6):371–379. [PubMed: 21184355]

Flagel SB, Clark JJ, Robinson TE, Mayo L, Czuj A, Willuhn I, et al. A selective role for dopamine in stimulus-reward learning. Nature. 2011; 469(7328):53–57. [PubMed: 21150898]

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

Foster DJ, Wilson MA. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature. 2006; 440(7084):680–683. [PubMed: 16474382]

Foster DJ, Wilson MA. Hippocampal theta sequences. Hippocampus. 2007; 17(11):1093–1099. [PubMed: 17663452]

Fowler H, Miller NE. FACILITATION AND INHIBITION OF RUNWAY PERFORMANCE BY HIND-AND FOREPAW SHOCK OF VARIOUS INTENSITIES. Journal of Comparative and Physiological Psychology. 1963; 56:801–805. [PubMed: 14050163]

Fudim OK. Sensory preconditioning of flavors with a formalin-produced sodium need. Journal of Experimental Psychology: Animal Behavior Processes. 1978; 4(3):276–285. [PubMed: 567670]

Ganesan R, Pearce JM. Effect of changing the unconditioned stimulus on appetitive blocking. Journal of Experimental Psychology: Animal Behavior Processes. 1988; 14(3):280–291. [PubMed: 3404082]

Gao Q, Horvath TL. Neuronal control of energy homeostasis. FEBS Lett. 2008; 582(1):132–141. [PubMed: 18061579]

Gershman SJ, Markman AB, Otto AR. Retrospective Revaluation in Sequential Decision Making: A Tale of Two Systems. Journal of Experimental Psychology: General. 2012

Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. #Neuron#. 2010; 66(4): 585–595. [PubMed: 20510862]

Goto Y, Grace AA. Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. Nature Neuroscience. 2005; 8(6):805–812.

Hall J, Parkinson JA, Connor TM, Dickinson A, Everitt BJ. Involvement of the central nucleus of the amygdala and nucleus accumbens core in mediating Pavlovian influences on instrumental behaviour. European Journal of Neuroscience. 2001; 13(10):1984–1992. [PubMed: 11403692]

Hassabis D, Kumaran D, Vann SD, Maguire EA. Patients with hippocampal amnesia cannot imagine new experiences. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(5):1726–1731. [PubMed: 17229836]

Herrnstein RJ. Levels of stimulus control: a functional approach. Cognition. 1990; 37(1–2):133–166. [PubMed: 2269005]

Hikosaka O, Nakahara H, Rand MK, Sakai K, Lu X, Nakamura K, et al. Parallel neural networks for learning sequential procedures. Trends in Neurosciences. 1999; 22(10):464–471. [PubMed: 10481194]

Hinton GE, Ghahramani Z. Generative models for discovering sparse distributed representations. Philos Trans R Soc Lond B Biol Sci. 1997; 352(1358):1177–1190. [PubMed: 9304685]

Holland PC. Conditioned stimulus as a determinant of the form of the Pavlovian conditioned response. Journal of Experimental Psychology: Animal Behavior Processes. 1977; 3(1):77–104. [PubMed: 845545]

Holland PC. Event representation in Pavlovian conditioning: image and action. Cognition. 1990; 37(1–2):105–131. [PubMed: 2269004]

Holland PC. Relations between Pavlovian-instrumental transfer and reinforcer devaluation. Journal of Experimental Psychology: Animal Behavior Processes. 2004; 30(2):104–117. [PubMed: 15078120]

Holland PC, Gallagher M. Double dissociation of the effects of lesions of basolateral and central amygdala on conditioned stimulus-potentiated feeding and Pavlovian-instrumental transfer. Eur J Neurosci. 2003; 17(8):1680–1694. [PubMed: 12752386]

Holland PC, Lasseter H, Agarwal I. Amount of training and cue-evoked taste-reactivity responding in reinforcer devaluation. J Exp Psychol Anim Behav Process. 2008; 34(1):119–132. [PubMed: 18248119]

Holland PC, Rescorla RA. The effect of two ways of devaluing the unconditioned stimulus after first- and second-order appetitive conditioning. Journal of Experimental Psychology: Animal Behavior Processes. 1975; 1(4):355. [PubMed: 1202141]

Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Computational Biology. 2012; 8(3):e1002410–e1002410. [PubMed: 22412360]

Hyman SE, Malenka RC, Nestler EJ. Neural mechanisms of addiction: the role of reward-related learning and memory. Annual Review of Neuroscience. 2006; 29:565–598.

Jenkins HM, Moore BR. The form of the auto-shaped response with food or water reinforcers. Journal of the Experimental Analysis of Behavior. 1973; 20(2):163–181. [PubMed: 4752087]

Johnson A, Redish AD. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. Journal of Neuroscience. 2007; 27(45):12176–12189. [PubMed: 17989284]

Jones JL, Esber GR, McDannald MA, Gruber AJ, Hernandez A, Mirenzi A, et al. Orbitofrontal cortex supports behavior and learning using inferred but not cached values. Science. 2012; 338(6109): 953–956. [PubMed: 23162000]

Keramati M, Dezfouli A, Piray P. Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Computational Biology. 2011; 7(5):e1002055–e1002055. [PubMed: 21637741]

Killcross S, Coutureau E. Coordination of actions and habits in the medial prefrontal cortex of rats. Cerebral Cortex. 2003; 13(4):400–408. [PubMed: 12631569]

Killcross S, Robbins TW, Everitt BJ. Different types of fear-conditioned behaviour mediated by separate nuclei within amygdala. Nature. 1997; 388(6640):377–380. [PubMed: 9237754]

Kishida KT, Sandberg SG, Lohrenz T, Comair YG, Saez I, Phillips PE, et al. Sub-second dopamine detection in human striatum. PLoS One. 2011; 6(8):e23291. [PubMed: 21829726]

Koob GF, Volkow ND. Neurocircuitry of addiction. Neuropsychopharmacology. 2010; 35(1):217–238. [PubMed: 19710631]

Korotkova TM, Brown RE, Sergeeva OA, Ponomarenko AA, Haas HL. Effects of arousal- and feeding-related neuropeptides on dopaminergic and GABAergic neurons in the ventral tegmental area of the rat. European Journal of Neuroscience. 2006; 23(10):2677–2685. [PubMed: 16817870]

Krause EG, Sakai RR. Richter and sodium appetite: from adrenalectomy to molecular biology. Appetite. 2007; 49(2):353–367. [PubMed: 17561308]

Krieckhaus EE, Wolf G. Acquisition of sodium by rats: interaction of innate mechanisms and latent learning. J Comp Physiol Psychol. 1968; 65(2):197–201. [PubMed: 5668303]

Lammel S, Lim BK, Malenka RC. Reward and aversion in a heterogeneous midbrain dopamine system. Neuropharmacology. 2013

Lammel S, Lim BK, Ran C, Huang KW, Betley MJ, Tye KM, et al. Input-specific control of reward and aversion in the ventral tegmental area. Nature. 2012; 491(7423):212–217. [PubMed: 23064228]

Leyton M, Vezina P. On cue: striatal ups and downs in addictions. Biol Psychiatry. 2012; 72(10):e21–22. [PubMed: 22789688]

Li J, Daw ND. Signals in human striatum are appropriate for policy update rather than value prediction. J Neurosci. 2011; 31(14):5504–5511. [PubMed: 21471387]

Lovibond PF. Appetitive Pavlovian-instrumental interactions: effects of inter-stimulus interval and baseline reinforcement conditions. Quarterly Journal of Experimental Psychology B, Comparative and Physiological Psychology. 1981; 33(Pt 4):257–269.

Lovibond PF. Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. Journal of Experimental Psychology: Animal Behavior Processes. 1983; 9(3):225–247. [PubMed: 6153052]

Mackintosh, NJ. Conditioning and Associative Learning. Oxford University Press; 1983.

Mahler SV, Berridge KC. What and when to "want"? Amygdala-based focusing of incentive salience upon sugar and sex. Psychopharmacology. 2012; 221(3):407–426. [PubMed: 22167254]

Marr, D. Vision. Freeman; 1982.

Matsumoto M, Hikosaka O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature. 2009; 459(7248):837–841. [PubMed: 19448610]

McAllister DE, McAllister WR, Hampton SR, Scoles MT. Escape-from-fear performance as affected by handling method and an additional CS-shock treatment. Animal Learning & Behavior. 1980; 8(3):417–423.

McClure SM, Daw ND, Montague PR. A computational substrate for incentive salience. Trends in Neurosciences. 2003; 26(8):423–428. [PubMed: 12900173]

McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G. Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. Journal of Neuroscience. 2011; 31(7):2700–2705. [PubMed: 21325538]

McDannald MA, Takahashi YK, Lopatina N, Pietras BW, Jones JL, Schoenbaum G. Model-based learning and the contribution of the orbitofrontal cortex to the model-free world. Eur J Neurosci. 2012; 35(7):991–996. [PubMed: 22487030]

Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience. 1996; 16(5):1936–1947. [PubMed: 8774460]

Moore RY, Bloom FE. Central catecholamine neuron systems: anatomy and physiology of the dopamine systems. Annu Rev Neurosci. 1978; 1:129–169. [PubMed: 756202]

Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H. Midbrain dopamine neurons encode decisions for future action. Nature Neuroscience. 2006; 9(8):1057–1063.

Morse WH, Mead RN, Kelleher RT. Modulation of elicited behavior by a fixed-interval schedule of electric shock presentation. Science. 1967; 157(3785):215–217. [PubMed: 17806273]

Murschall A, Hauber W. Inactivation of the ventral tegmental area abolished the general excitatory influence of Pavlovian cues on instrumental performance. Learning and Memory. 2006; 13(2): 123–126. [PubMed: 16547159]

Nieh EH, Kim SY, Namburi P, Tye KM. Optogenetic dissection of neural circuits underlying emotional valence and motivated behaviors. Brain Res. 2013; 1511:73–92. [PubMed: 23142759]

Niv Y, Daw ND, Joel D, Dayan P. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl). 2007; 191(3):507–520. [PubMed: 17031711]

O'Doherty JP. Contributions of the ventromedial prefrontal cortex to goal-directed action selection. Ann N Y Acad Sci. 2011; 1239:118–129. [PubMed: 22145881]

Pavlov, IP. Conditioned reflexes. Courier Dover Publications; 1927.

Pecina S, Berridge KC. Dopamine or opioid stimulation of nucleus accumbens similarly amplify cue-triggered 'wanting' for reward: Entire core and medial shell mapped as substrates for PIT enhancement. European Journal of Neuroscience. 2013 Online first.

Pezzulo G, Rigoli F, Chersi F. The Mixed Instrumental Controller: using Value of Information to combine habitual choice and mental simulation. Frontiers in Psychology. 2013; 4:92–92. [PubMed: 23459512]

Pfeiffer BE, Foster DJ. Hippocampal place-cell sequences depict future paths to remembered goals. Nature. 2013; 497(7447):74–79. [PubMed: 23594744]

Phillips PE, Stuber GD, Heien ML, Wightman RM, Carelli RM. Subsecond dopamine release promotes cocaine seeking. Nature. 2003; 422(6932):614–618. [PubMed: 12687000]

Prevost C, Liljeholm M, Tyszka JM, O'Doherty JP. Neural correlates of specific and general Pavlovian-to-Instrumental Transfer within human amygdalar subregions: a high-resolution fMRI study. J Neurosci. 2012; 32(24):8383–8390. [PubMed: 22699918]

Puterman ML. Markov decision processes: discrete stochastic dynamic programming. 2009; 414 Wiley.com.

Rescorla RA. Effect of US habituation following conditioning. J Comp Physiol Psychol. 1973; 82(1): 137–143. [PubMed: 4684968]

Rescorla RA. Effect of inflation of the unconditioned stimulus value following conditioning. Journal of Comparative and Physiological Psychology. 1974; 86(1):101.

Rescorla RA. Pavlovian conditioning. It's not what you think it is. Am Psychol. 1988; 43(3):151–160. [PubMed: 3364852]

Rescorla RA, Freberg L. Extinction of within-compound flavor associations. Learning and Motivation. 1978; 9(4):411–427.

Rescorla RA, Solomon RL. Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. Psychological Review. 1967; 74(3):151–182. [PubMed: 5342881]

Richard JM, Berridge KC. Nucleus Accumbens Dopamine/Glutamate Interaction Switches Modes to Generate Desire versus Dread: D1 Alone for Appetitive Eating But D1 and D2 Together for Fear. The Journal of Neuroscience. 2011; 31(36):12866–12879. [PubMed: 21900565]

Rizley RC, Rescorla RA. Associations in second-order conditioning and sensory preconditioning. Journal of Comparative and Physiological Psychology. 1972; 81(1):1–11. [PubMed: 4672573]

Robinson MJF, Berridge KC. Instant transformation of learned repulsion into motivational "wanting". Current Biology. 2013; 23(4):282–289. [PubMed: 23375893]

Robinson TE, Berridge KC. The neural basis of drug craving: an incentive-sensitization theory of addiction. Brain Res Brain Res Rev. 1993; 18(3):247–291. [PubMed: 8401595]

Robinson TE, Berridge KC. The incentive sensitization theory of addiction: some current issues. Philos Trans R Soc Lond B Biol Sci. 2008; 363(1507):3137–3146. [PubMed: 18640920]

Robinson TE, Kolb B. Structural plasticity associated with exposure to drugs of abuse. Neuropharmacology. 2004; 47:33–46. [PubMed: 15464124]

Roesch MR, Calu DJ, Schoenbaum G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nature Neuroscience. 2007; 10(12):1615–1624.

Roitman MF, Stuber GD, Phillips PE, Wightman RM, Carelli RM. Dopamine operates as a subsecond modulator of food seeking. J Neurosci. 2004; 24(6):1265–1271. [PubMed: 14960596]

Rossi MA, Sukharnikova T, Hayrapetyan VY, Yang L, Yin HH. Operant self-stimulation of dopamine neurons in the substantia nigra. PLoS One. 2013; 8(6):e65799. [PubMed: 23755282]

Rummery, G.; Niranjan, M. On-line Q-learning using connectionist systems. 1994.

Salamone JD, Correa M. Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. Behavioural Brain Research. 2002; 137(1–2):3–25. [PubMed: 12445713]

Samejima K, Ueda Y, Doya K, Kimura M. Representation of action-specific reward values in the striatum. Science. 2005; 310(5752):1337–1340. [PubMed: 16311337]

Saunders BT, Robinson TE. The role of dopamine in the accumbens core in the expression of Pavlovian-conditioned responses. The European Journal of Neuroscience. 2012; 36(4):2521–2532. [PubMed: 22780554]

Schacter DL, Addis DR, Buckner RL. Episodic simulation of future events: concepts, data, and applications. Annals of the New York Academy of Sciences. 2008; 1124:39–60. [PubMed: 18400923]

Schacter DL, Addis DR, Hassabis D, Martin VC, Spreng RN, Szpunar KK. The future of memory: remembering, imagining, and the brain. Neuron. 2012; 76(4):677–694. [PubMed: 23177955]

Schulkin J, Arnell P, Stellar E. Running to the taste of salt in mineralocorticoid-treated rats. Horm Behav. 1985; 19(4):413–425. [PubMed: 4085995]

Schultz W. Predictive reward signal of dopamine neurons. Journal of Neurophysiology. 1998; 80(1):1–27. [PubMed: 9658025]

Schultz W. Behavioral Theories and the Neurophysiology of Reward. Annu Rev Psychol. 2006; 57:87–115. [PubMed: 16318590]

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275(5306):1593–1599. [PubMed: 9054347]

Simon DA, Daw ND. Neural correlates of forward planning in a spatial decision task in humans. Journal of Neuroscience. 2011; 31(14):5526–5539. [PubMed: 21471389]

Smith KS, Berridge KC, Aldridge JW. Disentangling pleasure from incentive salience and learning signals in brain reward circuitry. Proc Natl Acad Sci U S A. 2011; 108(27):E255–264. [PubMed: 21670308]

Smith KS, Virkud A, Deisseroth K, Graybiel AM. Reversible online control of habitual behavior by optogenetic perturbation of medial prefrontal cortex. Proceedings Of The National Academy Of Sciences Of The United States Of America. 2012; 109(46):18932–18937. [PubMed: 23112197]

Spreng RN, Mar RA, Kim ASN. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative metaanalysis. Journal of Cognitive Neuroscience. 2009; 21(3):489–510. [PubMed: 18510452]

Stouffer EM, White NM. A latent cue preference based on sodium depletion in rats. Learning & Memory. 2005; 12(6):549–552. [PubMed: 16287723]

Suri RE, Schultz W. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience. 1999; 91(3):871–890. [PubMed: 10391468]

Sutton RS. Learning to predict by the methods of temporal differences. Machine Learning. 1988; 3(1):9–44.

Sutton, RS.; Barto, AG. Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning). The MIT Press; 1998.

Swanson LW. The projections of the ventral tegmental area and adjacent regions: a combined fluorescent retrograde tracer and immunofluorescence study in the rat. Brain Research Bulletin. 1982; 9(1–6):321–353. [PubMed: 6816390]

Talmi D, Seymour B, Dayan P, Dolan RJ. Human pavlovian-instrumental transfer. Journal of Neuroscience. 2008; 28(2):360–368. [PubMed: 18184778]

Thistlethwaite D. Conditions of irrelevant-incentive learning. Journal of Comparative and Physiological Psychology. 1952; 45(6):517. [PubMed: 13000023]

Thomas MJ, Kalivas PW, Shaham Y. Neuroplasticity in the mesolimbic dopamine system and cocaine addiction. Br J Pharmacol. 2008; 154(2):327–342. [PubMed: 18345022]

Timberlake W, Grant DL. Auto-shaping in rats to the presentation of another rat predicting food. Science. 1975; 190(4215):690–692.

Tindell AJ, Berridge KC, Zhang J, Peciña S, Aldridge JW. Ventral pallidal neurons code incentive motivation: amplification by mesolimbic sensitization and amphetamine. Eur J Neurosci. 2005; 22(10):2617–2634. [PubMed: 16307604]

Tindell AJ, Smith KS, Berridge KC, Aldridge JW. Dynamic Computation of Incentive Salience: “Wanting” What Was Never “Liked”. Journal of Neuroscience. 2009; 29(39):12220–12228. [PubMed: 19793980]

Toates, F. Motivational Systems. Cambridge: Cambridge University Press; 1986.

Tolman EC. Cognitive Maps in Rats and Men. The Psychological Review. 1948; 55:189–208.

Tolman EC. The nature and functioning of wants. Psychol Rev. 1949; 56(6):357–369. [PubMed: 15392594]

Tolman EC. Performance vectors and the unconscious. Acta Psychologica. 1955; 11:31–40.

Tomie A. Locating reward cue at response manipulandum (CAM) induces symptoms of drug abuse. Neuroscience and Biobehavioral Reviews. 1996; 20(3):31.

van der Meer MAA, Johnson A, Schmitzer-Torbert NC, Redish AD. Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. Neuron. 2010; 67(1):25–32. [PubMed: 20624589]

Vezina P, Leyton M. Conditioned cues and the expression of stimulant sensitization in animals and humans. Neuropharmacology. 2009; 56(Suppl 1):160–168. [PubMed: 18657553]

Volkow ND, Wang GJ, Fowler JS, Tomasi D. Addiction circuitry in the human brain. Annual Review of Pharmacology and Toxicology. 2012; 52:321–336.

Wanat MJ, Willuhn I, Clark JJ, Phillips PE. Phasic dopamine release in appetitive behaviors and drug addiction. Curr Drug Abuse Rev. 2009; 2(2):195–213. [PubMed: 19630749]

Wassum KM, Ostlund SB, Balleine BW, Maidment NT. Differential dependence of Pavlovian incentive motivation and instrumental incentive learning processes on dopamine signaling. Learning & memory. 2011; 18(7):475–483. [PubMed: 21693635]

Watkins, CJCH. Learning from Delayed Rewards. University of Cambridge; 1989.

Wilson CL, Sherman JE, Holman EW. Aversion to the reinforcer differentially affects conditioned reinforcement and instrumental responding. J Exp Psychol Anim Behav Process. 1981; 7(2):165–174. [PubMed: 7241052]

Wirsig CR, Grill HJ. Contribution of the rat’s neocortex to ingestive control: I. Latent learning for the taste of sodium chloride. Journal of comparative and physiological psychology. 1982; 96(4):615–615. [PubMed: 7119179]

Wise RA. Roles for nigrostriatal--not just mesocorticolimbic--dopamine in reward and addiction. Trends Neurosci. 2009; 32(10):517–524. [PubMed: 19758714]

Witten IB, Steinberg EE, Lee SY, Davidson TJ, Zalocusky KA, Brodsky M, et al. Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. Neuron. 2011; 72(5):721–733. [PubMed: 22153370]

Wolf ME, Ferrario CR. AMPA receptor plasticity in the nucleus accumbens after repeated exposure to cocaine. Neurosci Biobehav Rev. 2010; 35(2):185–211. [PubMed: 20109488]

Wunderlich K, Dayan P, Dolan RJ. Mapping value based planning and extensively trained choice in the human brain. Nature Neuroscience. 2012; 15(5):786–791.

Wyvell CL, Berridge KC. Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: enhancement of reward “wanting” without enhanced “liking” or response reinforcement. Journal of Neuroscience. 2000; 20(21):8122–8130. [PubMed: 11050134]

Yager LM, Robinson TE. A classically conditioned cocaine cue acquires greater control over motivated behavior in rats prone to attribute incentive salience to a food cue. Psychopharmacology. 2013; 226(2):217–228. [PubMed: 23093382]

Zaghloul KA, Blanco JA, Weidemann CT, McGill K, Jaggi JL, Baltuch GH, et al. Human substantia nigra neurons encode unexpected financial rewards. Science. 2009; 323(5920):1496–1499. [PubMed: 19286561]

Zener K, McCurdy HG. Analysis of Motivational Factors in Conditioned Behavior: I. The Differential Effect of Changes in Hunger Upon Conditioned, Unconditioned, and Spontaneous Salivary Secretion. Journal of Psychology. 1939; 8(2):321–350.

Zhang J, Berridge KC, Tindell AJ, Smith KS, Aldridge JW. A neural computational model of incentive salience. PLoS Comput Biol. 2009; 5(7):e1000437. [PubMed: 19609350]

Zigman JM, Jones JE, Lee CE, Saper CB, Elmquist JK. Expression of ghrelin receptor mRNA in the rat and the mouse brain. Journal of Comparative Neurology. 2006; 494(3):528–548. [PubMed: 16320257]

Appendix: Model-free and model-based computations

In this appendix, we provide a very brief description of two classes of computational approach to instant Pavlovian revaluation. One is the algorithmic suggestion from Zhang and colleagues (Zhang et al., 2009); the other comes from the computational framework of reinforcement learning (Sutton, 1988).

Both methods start from the observation that the value $V(s_t, m)$ of a circumstance s_t, which is signalled by a CS, under a motivational state m is intended to be the expected long term, discounted, future utility available starting from that circumstance:

$$V(s_t,m)=\mathbb{E}\left[\sum\nolimits_{\tau=0}\gamma^{\tau}r(s_{t+\tau},m)\right]$$

where γ is a temporal discount factor and $r(s, m)$ is the net *utility* in state m of any UCS that is provided under circumstance s. If we only consider a single state m, and write $r_t = r(s_t, m)$ and drop the expectation for simplicity, then we have

$$V(s_t,m)=r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\cdots\triangleq r_t+\gamma V(s_{t+1},m) \quad \text{(A1)}$$

where the last equality holds if the prediction at the next timestep is correct. This is a consistency condition. Model-free temporal difference learning (Sutton, 1988) uses the discrepancy between right and left sides of this equation

$$\delta_t=r_t+\gamma V(s_{t+1},m)-V(s_t,m)$$

as a prediction error signal to criticise the original estimate, and so specifies

$$V(s_t,m) \leftarrow V(s_t,m)+\varepsilon\delta_t \quad \text{(A2)}$$

where ε is a learning rate. Albeit in the absence of consideration of revaluation experiments, (McClure et al., 2003) suggested that δ_t has the correct properties to act as an incentive salience signal. Importantly, the quantities r_t, s_t, s_{t+1} that determine δ_t are all available directly from the input; this is why temporal difference learning can proceed without a model of the world. However, it is apparent that δ_t is tied to a particular motivational state. If the state changes from m to $\tilde{m}$, the model-free learning mechanism provides no hint as to how to change $V(s_t, m)$ to $V(s_t, \tilde{m})$

The kappa transformation of prediction error cache into incentive salience

The kappa transformation accomplishes an instant revaluation of a Pavlovian CS (Zhang et al., 2009). It is based on modulating a particular sort of cached model-free prediction by a factor called κ. According to one version of this, the incentive salience value, $\tilde{V}(s_t, \tilde{m})$ of a Pavlovian CS that defines the circumstance s_t at the moment of re-encounter is instantly adapted from equation (A1) to

$$\tilde{V}(s_t,\tilde{m})=r_t * \kappa+\gamma V(s_{t+1},m) \quad \text{(A3)}$$

where κ is a dynamic physiological factor that provides a phenomenological model of the effect of the new state $\tilde{m}$ of the brain and body associated with that outcome. Given suitable values for the various factors, it is possible to explain the results for experiments that exhibit instant revaluation.

Model-based learning and representations of outcome

There are two key characteristics of the model-free learning rule of equation (A2). First, it is purely written in terms of utilities or estimates of sums of those utilities, and so retains no information about UCS identities that underlie them. This is its central model-free characteristic. Second, it ensures that the prediction is of the full, long-term, reward consistent with starting from that circumstance. This is why the model-free system is formally no less patient than the model-based system.

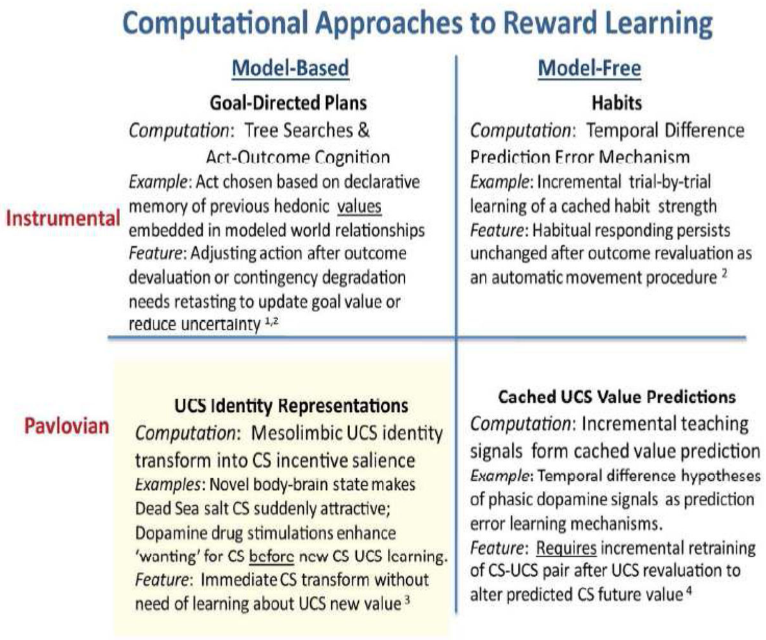
Equation (A3) is implicitly at variance from both these characteristics. First, in order for the κ transformation to operate correctly, it is essential that the *identity* of the reward that is immediately expected be known, so that its appropriate κ can be applied. This is more information than mere utility. In fact it is necessary for $\kappa(m, \tilde{m})$ to be a function of both the original m and current $\tilde{m}$ motivational states for the value transformation to be correct. Second, it is necessary to be able to split off the immediate expected utility r_t from the full, model-free, prediction $V(s_t, m)$ of equation (A1), and also to be able to know which next circumstance s_{t+1} to expect, in order to compute the $\gamma V(s_{t+1}, m)$ component in the equation. That is, it is necessary to know the function $r(s, m)$ and the transition function $T_{xy} = P(s_{t+1} = y|s_t = x)$ (where the latter is appropriate to a dynamic structure called a Markov chain (Puterman, 2009), which puts the stochasticity back into equation (A1)), which indicates how likely each successor state *y* is starting from state *x*. These exactly constitute a model of the domain.

If one had such a model, one could write equation (A1) in the new state $\tilde{m}$ exactly as

$$V(s_t,\tilde{m})=r(s_t,\tilde{m})+\sum\nolimits_{s_{t+1}} T_{s_t s_{t+1}} r(s_{t+1},\tilde{m})+\sum\nolimits_{s_{t+2}} T^2_{s_t s_{t+1}} r(s_{t+2},\tilde{m})+\cdots \quad \text{(A4)}$$

The sub-field of model-based reinforcement learning (Sutton & Barto, 1998) provides a number of methods for evaluating equation (A4) explicitly, for instance by direct construction or imagination of the sum over future possibilities for s_{t+1} and s_{t+2}, via $T_{s_t s_t+1}$ and $T^2_{s_t s_t+1}$, etc. It is because the model can use a current estimate of the transition model $T_{s_t s_t+1}$ that it can report correctly the effect of changes in the transition contingencies in the environment. Equally, it is because the model knows the identity of the outcomes associated with the circumstances $s_{t+\tau}$ that it can report the utilities $r(s_{t+\tau}, \tilde{m})$ according to the current motivational state (as in the Dead Sea salt experiment), provided that cortical structures such as the orbitofrontal cortex or subcortical mechanisms do not need learning in state $\tilde{m}$ (either by retasting or instrumental incentive learning) for such reports to be accurate.

Equation (A3) can be seen as an heuristic method for calculating an approximate modelbased prediction in which model-free values are used to replace the utility contributions from all imagined circumstances beyond s_{t+1}. This heuristic leads to a method that lies somewhere between model-free and model-based, and is essential for searching substantial-sized domains (indeed being common in planning in games such as chess; (Campbell, Hoane Jr, & Hsu, 2002)).



Computational Approaches to Reward Learning

	Model-Based	Model-Free
Instrumental	**Goal-Directed Plans** *Computation*: Tree Searches & Act-Outcome Cognition *Example*: Act chosen based on declarative memory of previous hedonic values embedded in modeled world relationships *Feature*: Adjusting action after outcome devaluation or contingency degradation needs retasting to update goal value or reduce uncertainty [1,2]	**Habits** *Computation*: Temporal Difference Prediction Error Mechanism *Example*: Incremental trial-by-trial learning of a cached habit strength *Feature*: Habitual responding persists unchanged after outcome revaluation as an automatic movement procedure [2]
Pavlovian	**UCS Identity Representations** *Computation*: Mesolimbic UCS identity transform into CS incentive salience *Examples*: Novel body-brain state makes Dead Sea salt CS suddenly attractive; Dopamine drug stimulations enhance 'wanting' for CS before new CS UCS learning. *Feature*: Immediate CS transform without need of learning about UCS new value [3]	**Cached UCS Value Predictions** *Computation*: Incremental teaching signals form cached value prediction *Example*: Temporal difference hypotheses of phasic dopamine signals as prediction error learning mechanisms. *Feature*: Requires incremental retraining of CS-UCS pair after UCS revaluation to alter predicted CS future value [4]

Figure 1.

A summary comparison of computational approaches to reward learning. Columns distinguish the two chief approaches in the computational literature: model-based versus model-free. Rows show the potential application of those approaches to Instrumental versus Pavlovian forms of reward learning (or equivalently to punishment or threat learning). We suggest the Pavlovian model-based cell (colored at lower left) has hitherto been comparatively neglected, as computational approaches have tended to treat Pavlovian learning as being purely model-free. However, evidence indicates that model-based Pavlovian learning happens and is used for mesolimbic-mediated instant transformations of motivation value. By contrast, instrumental model-based systems that model the value of an outcome, based on memory of its hedonic experience, may need to retaste or re-experience outcome again after revaluation in order to update model (see text for discussion and alternatives). Each cell contains a) a brief description of its characteristic computation, b) an example of behavioral or neural demonstrations in the experimental literature, and c) a distinguishing feature by which it can be recognized in behavioral or neural experimental findings. Citations: 1) (Dickinson & Balleine, 2010); 2) (Daw et al., 2005); 3) (Robinson & Berridge, 2013); 4) (Schultz et al., 1997).

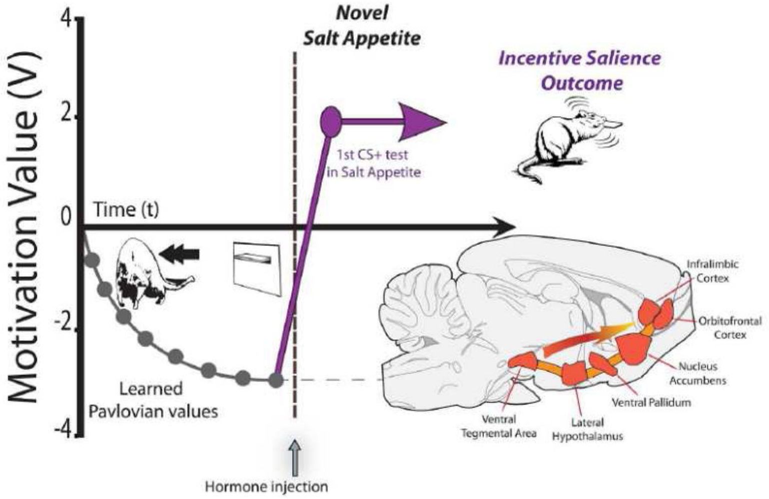


Figure 2.

Instant transformation of CS incentive salience observed in Dead sea salt study (Robinson & Berridge, 2013). Initial aversive Pavlovian training of CS+ with disgusting UCS taste produces gradual learned repulsion. CS+ value declines negatively over successive CS+ pairings with NaCl UCS (learned Pavlovian values). After training, sudden hormone injections induce novel state of salt appetite. CS value is transformed instantly into positive on very first next re-encounter in new appetite state (CS+ presented alone in crucial test, without salty UCS being retasted). Behaviorally, rats approach and nibble the CS+ lever that was previously associated with disgusting NaCl taste as UCS as avidly as a different CS previously associated with a pleasant sucrose UCS. Neurobiologically, mesolimbic brain activations were observed during combination of CS+ re-encounter plus novel appetite state in dopamine-related structures: ventral tegmentum, nucleus accumbens, prefrontal cortex, etc. Quantitative transformation depicted is based on (Zhang et al., 2009)'s computational model of incentive salience. Figure modified from (Robinson & Berridge, 2013) and (Zhang et al., 2009).