Rewards are Categories.

Erik J. Peterson
Dept. of Psychology
Colorado State University
Fort Collins, CO

The neural mechanisms of reinforcement learning are becoming increasingly clear following years of exciting and intense inquiry. However, due to their reliance on primary and secondary reward concepts, reinforcement learning theories can't account for two related facts. One, rewarding effects are observed in the absence of primary and secondary reinforcers (e.g., novelty, information and fictive outcomes). Two, value can be transferred by inference; no pairing is needed (e.g., stimulus generalization, optimistic firing). These atypical, or "cognitive rewards", have received little direct investigation; this thesis examines then a proposed mechanism that could underlie both these facts – by treating and modeling rewards as a kind of category, reward knowledge can be constructed and transferred (by similarity-based inference) to new situations. Using behavioral, fMRI, and computational data, this proposal was tested. Participants completed a stimulus-response task where classical rewards (e.g.,"Correct!" or "Win $1.") were replaced with pre-trained perceptual categories, one reward category for gains and one for losses. The reward for each trial was a unique, never before or again experienced, examplar from one of the two reward categories, distinguishing this task from higher-order conditioning paradigms were the same stimulus is repeatedly paired or presented. In total, the behavioral and neural data strongly suggest that cognitive rewards are in fact categories, categories which do substantively impact fMRI-based reinforcement learning signals in the brains of the human participants. It is then further argued that as category representations are a complete mechanistic explanation for the well established generalization of (classical) secondary reinforcers, rewards are categories – which represents a substantial change in how rewards are conceived, and modeled: the primary, to secondary, to higher-order conditioning paradigm is incomplete, perhaps even incorrect.

Contents

Introduction **4**
Classics, Expectations, and Tissues 6
Bad Prediction, No Cookie 15
Thinking About Thinking Rewarding Thoughts 19
Generally Generalizable. 22
Why Are We Here Again? 26

Chapter 2 – Task and Models **29**
On Task 29
3 Models and 2 Codes 38

Chapter 3 – fMRI analyses **51**
An Acquisition 51
Mobs of Blobs 54
Regions and Models 55
Model Results 68

Discussion **91**
Taking Us to Can 91
Are, Reflected in Error(s) 93
The Big Conclusion 98
Future Work 99

References **101**

Introduction

Birds will peck repeatedly, as mice will push levers, monkeys will hit buttons, and men will buy flowers, if each of these actions is followed by a primary reward – food, drink or sex. Buttons, levers and flowers have no value alone, so reinforcement theory goes, it is only through the *statistically regular pairing* with primary rewards that value is transferred (Rescorla, 1988). This is the classical view, and it has, in general, held up many years now (Iversen & Iversen, 2007). And indeed, the neural mechanisms of reinforcement learning are becoming increasingly clear following years of exciting and intense inquiry (for reviews see, Dayan and Daw (2008); Dayan and Niv (2008); Montague, King-Casas, and Cohen (2006)). However, due to their reliance on primary and secondary reward concepts, reinforcement learning theories can't account for two related facts. One, rewarding effects are observed in the absence of primary and secondary rewards (Hayden, Pearson, & Platt, 2009; Lohrenz, McCabe, Camerer, & Montague, 2007; Tricomi & Fiez, 2008; Jimura, Locke, & Braver, 2010). Two, value can be transferred by inference; no pairing is needed (Bromberg-Martin, Matsumoto, Hong, & Hikosaka, 2010; Hampton, Bossaerts, & O'Doherty, 2006). These two facts, it will be argued, are irreconcilable with traditional secondary or higher-order conditioning, and so require a new nomenclature. Primary, secondary and other higher-order rewards, will be denoted as "classical rewards", while other kinds of rewarding activities will be labeled as a "cognitive rewards" (dropping the quotes for convenience). The difference though is more than semantic. Cognitive rewards may represent a new conceptualization of rewards.

Using fMRI, behavioral, and computational data, the efficacy of one model system for cognitive rewards was examined, as was one possible computational mechanism for the creation and use of cognitive rewards: by treating and modeling cognitive rewards as a kind of category, arbitrary kinds of reward knowledge can be combined and transferred (by similarity-based inference) to new situations, thus accounting for two facts above.

This introduction has five parts. First is a review classical rewards and the reward prediction error account of dopamine function, along with its anatomical basis. Second, a critique of the rewards prediction error account is offered. Third is the case for cognitive rewards. Fourth is an argument for the logical and empirical necessity of generalizable reward representations, covering stimulus generalization and categorization along the way. In the fifth and final section, the exact goals and methods of this work are laid out.

Much time is spent in the first three parts reviewing reward-driven learning, particularly the reward prediction error hypothesis of phasic activity in the ventral tegmental area/substantia nigra pars compacta (VTA/SNc) and its neural substrates. This time is warranted as all computational models considered herein are viable and interesting only if the reward prediction error hypothesis is true. That is, the primary question of interest here, "Are rewards categories?", is asked through the lens of the reward prediction error hypothesis. Additionally, many of the neural regions of interest are of interest as they play key roles in reward prediction and reward prediction error calculations. For convenience, the phasic dopaminergic signal (i.e., the reward prediction error signal) from VTA/SNc to basal ganglia and cortical

areas will often be referred to as just "dopamine" or "dopaminergic activity". There are of course other kinds and roles of dopamine release (e.g., tonic activity driven by VTA/SNc (Schultz, 2007) or dendritic back-propagation in cortex (Jay, 2003)), however none of these are of immediate interest here.

Classics, Expectations, and Tissues

A pleasurable start. Classically rewards and reinforcers were linked to or operationally defined as food, water (O'Doherty, Buchanan, Seymour, & Dolan, 2006; Schultz, 2007) and sex though for, err, logistical reasons this is less often used in the laboratory. Classic rewards are certainly potent, having been used for over 50 successful years to study learning in animal models (Iversen & Iversen, 2007) and people (H. Kim, Shimojo, & O'Doherty, 2010; Montague et al., 2006). In the 1950's the first clue how rewards cause reinforcement arose in the electrical self-stimulation studies of Olds & Milner (1956), and, Crow (1972). Olds and colleges observed that when electrodes were placed in the dopaminergic midbrain, animals would vigorously and repeatedly self-stimulate by pressing the available button. By the 1970s Old's shocking work, combined with data from pharmacological studies of rats, electrochemical recordings, knowledge of the signaling mechanisms of dopamine receptors, as well as neuroleptic drug actions in Schizophrenic patients, lead to the first major theoretical proposal for dopamine's cognitive function - a signal for pleasure, sometimes called the anhedonia hypothesis (Wise, 1978). Within 10 years however it became clear that dopamine's role extends beyond signaling primary rewards and pleasure. Activity was seen following secondary rewards, novelty, salience, and in

other manipulations (Spanagel & Weiss, 1999; Salamone, Correa, Mingote, & Weber, 2005; Bromberg-Martin, Matsumoto, & Hikosaka, 2010b). More importantly, dopamine depleted animals continued to enjoy rewards, i.e., they still developed taste preferences, enhanced response vigor (Cannon & Palmiter, 2003) and continued to respond to opiates (Hnasko, Sotak, & Palmiter, 2005). In 1994 there was a surprise that would eventually account for many of anhedonia's deficits. Mirenowicz and Schultz (1994) reported that dopamingeric firing depended on how expected a reward was; this observation blossomed into the reward prediction error theory under review here.

Expectations matter. Continuing to work in Macaque Hollerman and Schultz (1998) more fully explored Mirenowicz and Schultz's (1994) observation, showing that unexpected rewards lead to increases in dopamine neurons' firing rates, fully expected rewards elicit no response, and expected rewards that fail to arrive lead to a dip in the baseline firing rate. Roesch, Calu, and Schoenbaum (2007) found the same patterns in rats while O'Doherty, Dayan, Friston, Critchley, and Dolan (2003) found them too in fMRI studies of the human striatum. Striatal BOLD changes reflect phasic dopamine activity (Schonberg et al., 2009; Surmeier, Ding, Day, Wang, & Shen, 2007)[1] . In ground breaking work, Waelti, Dickinson, and Schultz (2001) showed that stimulus-response learning and the dopaminergic signal are maximized not by reward reliability, but instead when rewards were intermittent (i.e., more rewards are not always better). Waelti et al. (2001), along with, Fiorillo, Tobler,

[1]Though this has recently come under question in rat models comparing single unit and field recordings to high resolution blood flow changes (Mishra et al., 2011)

and Schultz (2003) and Bayer and Glimcher (2005), successfully modeled changes in reward expectancy with a reward prediction error term derived from a reinforcement learning model[2] fit to each animal's behavior. The reward prediction error from the model strongly correlated with the dopamine response, both its increases and decreases. However expectancy related changes are not the only important prediction reinforcement learning models make for dopaminergic activity.

Value must transfer. If an initially neutral cue reliably predicts a reward the reinforcement learning equations require the prediction error (and thus the dopaminergic response) to transfer to the cue, thus mimicking Pavlovian conditioning. This behavior was observed in the dopamine response as well (Roesch et al., 2007; McClure, Berns, & Montague, 2003). In sum, all characteristic predictions made by the reinforcement learning equations[3] have been observed in the dopaminergic signal.

Outside of correlational evidence, reinforcement learning models are also statistically predictive of non-human animal's choice behaviors (Hampton & O'Doherty, 2007). Single doses of dopamine antagonists and agonists have demonstrated a casual relationship between dopamine levels and learning rate (Pizzagalli et al., 2008; Diaconescu, Menon, Jensen, Kapur, & McIntosh, 2010), which is broadly, though not exclusively, consistent with a reinforcement theory interpretation. Reward prediction terms have also been shown to statistically mediate cortical-striatal coupling (Ouden, Daunizeau, Roiser, Friston, & Stephan, 2010). Not limited to the above pre-

[2]Specifically the Rescorla-Wagner model, more on that later.

[3]Or to be precise, the Rescorla-Wagner and Temporal Difference family of reinforcement learning models

dictions and confirmatory findings, the reinforcement learning account has extended substantially, both theoretically and empirically.

Based on novel findings about novelty (Bunzeck & Düzel, 2006; Blatter & Schultz, 2006; Guitart-Masip, Bunzeck, Stephan, Dolan, & Duzel, 2010b) the reward prediction hypothesis has been extended to incorporate activity observed following presentation of novel stimuli (Kakade & Dayan, 2002) as well as to explain reward anticipatory firing via an average reward prediction error (Knutson & Wimmer, 2007). Another variation allowed for simultaneous neural implementations of model-free and model-based reinforcement learning (A. Smith, Li, Becker, & Kapur, 2006; Daw, Gershman, Seymour, Dayan, & Dolan, 2011). Alternative, but reconcilable, reinforcement learning equations have also been offered that allow for dissociation of first and second order conditioning, as well as explain Pavlovian to instrumental transfer (O'Reilly, Frank, Hazy, & Watz, 2007). The reward prediction hypothesis has also been incorporated into theoretical accounts of addiction (Redish, 2004) and been used to predict the salience of upcoming stimuli (Behrens, Woolrich, Walton, & Rushworth, 2007).

There are however several aspects of dopamine function which are, as yet, are unaccounted for theoretically. Matsumoto and Hikosaka (2009), reported a very broad set of dopaminergic firing patterns. In the classical (bivalent) view dopamine neurons should fire more for unexpected rewards or omission of punishment, less for reward omission and in response to aversive events. Instead H. Kim, Shimojo, and O'Doherty (2006); Matsumoto and Hikosaka (2009), found that some neurons respond bivalently but many others showed an inverse coding scheme, responding posi-

tively to aversive stimuli increasing as the punishment grew larger than expected, and decreasing as it grew smaller than expected. Yet other neurons responded positively to *both* appetitive and aversive conditions. In a separate experiment K. S. Smith, Berridge, and Aldridge (2011) demonstrated *simultaneous yet separate* tunings to reward value, reward expectancy, salience as well as to novelty. The dopamine response also appears to adaptively scale with past reward magnitudes. These dynamic range adjustments appeared similar to the reward value divided by the cumulative variance (Tobler, Fiorillo, & Schultz, 2005). If the reports above hold up, i.e., the dopaminergic response is complex, the bivalent view needs substantial refinement, as do perhaps our analysis techniques; many different models or neural coding schemes may *correctly* fit the same data, an issue which has received some prior attention outside of neuroimaging[4] (Chamberlin, 1965).

Networked plausibility. The dopaminergic firing patterns outlined above originate in the VTA/SNc, a small brainstem nucleus whose neurons project strongly to both the striatum, prefrontal cortex (PFC) and the hippocampus. Electrophysiological recordings of VTA/SNc neurons show two firing modes – tonic and phasic (for a review see Schultz (2007)). The phasic mode is of interest here, as it is this that reflects the reward predictions error. A reward prediction *error* signal of course requires a prediction, in this case a prediction future reward value and probability. How exactly such predictions are made is only partly understood. Candidate regions for this calculation include the striatum, the limbic system routed through the habenula, as well at the orbital, ventral medial and the dorsal lateral frontal cortices.

[4]For a (first) take on addressing this problem in fMRI, see p59).

Each is presented in turn, leaving the totality largely unintegrated, thus accurately reflecting the literature's state.

Selecting striatum. The striatum is the input area of the basal ganglia, a brain region involved in categorization, logical inference, habit formation, working memory and feedback mediated stimulus-response learning (Frank, Loughry, & O'Reilly, 2001; Jin & Costa, 2010; Schmitzer-Torbert & Redish, 2004; Seger, 2008; Seger & Miller, 2010; Yin & Knowlton, 2006). In stimulus-response learning, two of the four striatal subregions (the head of the caudate and the ventral striatum) process reward information (Yin, Ostlund, Knowlton, & Balleine, 2005; Yin, Ostlund, & Balleine, 2008; Schonberg et al., 2009). These two are highly innervated by projections from the VTA/SNc, but only the ventral striatum correlates with the reward prediction error signal (Haruno & Kawato, 2006; Seger & Miller, 2010). The remaining two regions (the body and tail of the caudate and the putamen) are involved with stimulus-response pair formation, specifically visual categorization and response selection, respectively (Seger, 2008; Seger & Miller, 2010). Though these regions also receive VTA/SNc projections and are sometimes sensitive to reward level (Bischoff-Grethe, Hazeltine, Bergren, Ivry, & Grafton, 2009) the BOLD signal does not correlate with the reward prediction error (Seger & Miller, 2010). Dopamine's exact role in these areas is less clear. In addition to projecting to basal ganglia, VTA/SNc also receives input from the internal globus pallidus or GPi (a major output structure of the basal ganglia with widespread cortical connections), thalamus, and the central nucleus of the amygdala (Botvinick, Niv, & Barto, 2008). Thus the striatum may form a value evaluation *and* action selecting loop, similar to the classic actor-critic reinforcement

learning architecture (Bornstein & Daw, 2011; Ito & Doya, 2011) though this view has recently received some strong criticisms (Joel, Niv, & Ruppin, 2002).

Intact dopamine projections and striatal function is necessary for rapid stimulus-response learning. Administering dopamine antagonists to human and non-human animals adversely affects stimulus-response learning (Pizzagalli et al., n.d.), as does lesioning the VTA/SNc. Complete lesions of the striatum also prevent stimulus-response learning (Packard & Knowlton, 2002). Administering dopamine agonists or the readily converted precursor L-DOPA leads to increases in response vigor and the ability of a Pavlovian-conditioned stimulus to bias unrelated instrumental responses (i.e., Pavlovian instrumental transfer) (Winterbauer & Balleine, 2007). Both Pavlovian instrumental transfer and response vigor are, in part, facilitated by phasic dopamine increasing activity in the ventral striatum. Unmedicated Parkinson's patients, who have low striatal dopamine levels, show marked decreases in stimulus-response learning with rewarding outcomes when compared to patients on medication and healthy age and intellect matched controls (Pizzagalli et al., n.d.). These same patients show an enhanced capability to learn from negative feedback which suggests that decreases in dopamine convey negative outcome information (Frank, Seeberger, & O'Reilly, 2004). Finally, there is a solid body of evidence suggesting that phasic dopamine alters the plasticity of neurons in the striatum which presumably facilitates stimulus-response learning (Calabresi, Picconi, Tozzi, & Filippo, 2007).

Linking with the limbic. Recent work has highlighted the habenula (a small nucleus posterior to the thalamus) as being especially important in generating reward

prediction error-like phasic activity in VTA/SNc. Besides its limbic connections (to the amygdala, hippocampus and the serotonergic dorsal raphe nucleus (Hikosaka, Sesack, Lecourtier, & Shepard, 2008)) the lateral habenula has reciprocal connections with the GPi and projects to the VTA/SNc. Based on this anatomy, it has been suggested that the habenula could serve a point of intersection between the striatum and the limbic system. Inline with this proposed role, the habenula can tonically inhibit or disinhibit dopamine release in VTA/SNc neurons, making VTA/SNc activity inversely correlated with the habenula. As habenula activity decreases, burst firing in VTA/SNc increases. Likewise as habenula firing increases, VTA/SNc activity temporarily pauses. Reversible chemical inhibition of habenula also increases VTA/SNc phasic activity (Hikosaka et al., 2008). Lesions to the habenula result in marked increases in dopamine levels in the dorsal and ventral striatum (Bromberg-Martin, Matsumoto, & Hikosaka, 2010a). Detailed dual recording studies of both areas *hint* that, combined, these areas calculate the value of stimulus-response pairs (Bromberg-Martin, Matsumoto, & Hikosaka, 2010a). In summary, the GPi, with its access to cortical inputs via the striatum, and habenula, with its capability for altering VTA/SNc activity, may form the physiological loop necessary to calculate of the reward prediction error.

Front and center, lateral too. Orbital frontal cortex has been repeatedly shown in neuroimaging (O'Doherty, Kringelbach, Rolls, Hornak, & Andrews, 2001) and lesion (Hornak et al., 2004) studies to encode the absolute value of rewarding or punishing outcomes, thus playing a pivotal role in the new field of neuroeconomics (Glimcher, Dorris, & Bayer, 2005). However orbital frontal areas are more than a

simple value store. They also play a role in response and outcome recall, responses selection (Rudebeck et al., 2008; Furuyashiki, Holland, & Gallagher, 2008), motivation, pain and pleasure (Atlas, Bolger, Lindquist, & Wager, 2010), outcome anticipation and prediction (Tanaka et al., 2006; Roesch & Olson, 2007) and causal attribution (Tanaka, Balleine, & O'Doherty, 2008). Additionally, orbital function is highly dependent on the basolateral amygdala (O'Doherty, 2003) offering a path for orbital activity to inform reward prediction error calculation.

Besides value, estimating the likelihood a reward will occur is the other key reward prediction calculation. Correlations with the chance of receiving a reward (Tobler, Christopoulos, O'Doherty, Dolan, & Schultz, 2009), the variance of expected value (Kahnt, Heinzle, Park, & Haynes, 2010) as well as with risk-seeking behaviors (Tobler, O'Doherty, Dolan, & Schultz, 2007) have all been reported in dorsolateral prefrontal areas. These same areas have also been implicated in inter-temporal choice, i.e., deciding between immediate and delayed rewards (S. Kim, Hwang, Seo, & Lee, 2009; S. Kim, Hwang, & Lee, 2008).

Unlike orbital and dorsolateral cortices, the nearby ventromedial prefrontal cortex encodes information on both reward magnitude and reward probability (Knutson et al, 2005), suggesting this region may synthesize orbital and dorsolateral activities. Unlike orbital cortex, whose activity is linked only to stimuli, value in the ventromedial PFC is associated with actions (Glascher, Hampton & ODoherty 2009). Furthering a possible integrative interpretation of dorsolateral activities, values in this area seem to be embedded in a "higher order" or abstract representation that reflects the statistical structure of the underlying task. This abstract representation seems

to support the inference of value (Hampton et al., 2006). However ventromedial cortex's role in inference may be more general. It also plays a role in inferring the valued expectations of an opponent in two-player game of strategy (Hampton, Bossaerts & O'Doherty, 2008) as well as inferring others intentions outside of economic games (Cooper, Kreps, Wiebe, Pirkl & Knutson, 2010).

Bad Prediction, No Cookie

The hypothesis that rewards are represented as categories in the brain assumes that dopamine, specifically that phasic activity in VTA/SNc leads to brief elevations in the concentration of dopamine in striatal and cortical areas, acts to stamp in stimulus-response relationships. What follows is a critical look at that assumption.

Not cortex, colliculus. Recent work by Dommett et al. (2005) is a *potentially* deadly issue for the reward prediction hypothesis. The reward prediction theory requires very specific timing in order to map rewards to states and actions. This requirement is satisfied by dopamine activity patterns, which are typically seen as 100 ms long burst about 100 ms after the initial stimulus. Dommett et al (2005) were however concerned that 100 ms may not be sufficient time for a saccade and detailed visual processing, followed by prefrontal examination/valuation, and finally reward prediction error calculation (the minimum set of steps likely required for reward prediction error formulation, see p13 for more). To examine this concern, Dommet et al (2005) disinhibited neurons in the brains of anesthetized rats in both the superior

colliculus, a brainstem visual processing area, and early visual cortex with a GABA antagonist (which temporarily restores neural activity in the normally unresponsive neurons of an anesthetized animal) then exposed the animals opened eyes to a set of 2 Hz light pulses while recording dopamine cells in the SNc/VTA as well as neurons in both visual cortex and the superior colliculus. Disinhibition of the visual cortex lead to no changes in dopamine firing. Superior colliculus disinhibition caused about half the recorded VTA/SNc neurons to fire, firing which was similar in character to that typically observed following an unexpected rewarding event. Another third of the dopamine cells displayed a pause in activity, similar to that observed when an expected reward fails to arrive (Mirenowicz & Schultz, 1994). The remaining cells responded first positively then negatively. From this the authors concluded that the superior colliculus and not the visual cortex is an effective activator of VTA/SNC neurons[5]. This is a problem as the superior colliculus responds only to very limited range of visual stimuli – appearance, disappearance, or movement of objects as well as luminance changes. It does not respond to contrast, velocity, wavelength or the geometric configuration of stimuli (Dommett et al., 2005). That is, the superior colliculus couldn't realistically extract sufficient information from the visual stimuli used in nearly all the studies of reward to date. Instead, Dommett et al (2005) argue that "[all of the many reward studies] can be solved based on luminance changes and/or of the position of specific reward-related visual stimuli". In other words, the expectancy-of-reward related changes in phasic dopamine are an artifact of the task

[5]Another reason for more cautious interpretation: Dommet et al's (2005) disinhibition experiments in visual cortex were very limited (N=4 cells compared to 30 for the superior colliculus experiments). Nor did they assess the extent of disinhibition in visual cortex.

design.

Dommet et al's (2005) interpretation is however too strong. At best they have shown that there is not substantive direct connection between visual areas and the VTA/SNc. If there was only a single downstream synapse between visual cortex and SNc/VTA their protocol would not have disinhibited it and so would have failed to elicit a dopamine response during visual stimulation. Given the brain's high degree of inter-connectivity and probable small-world architecture (Bassett & Bullmore, 2006), a failure to find a direct anatomical relation is not in and of itself conclusive. Additionally, Dommett et al's (2005) concern may also be inconsistent with EEG studies of human cognition (?, ?). Objections aside, fully establishing the temporal plausibility of the dopamine signal is still an open and very important question.

They want, I want. Standing in opposition to both the anhedonia and reward prediction hypotheses is the incentive salience account, which is derived from the brute fact that addicts often greatly want drugs of abuse, but once drugs are received addicts, or their animal model counterparts, do not report an excess of pleasure or liking (Robinson & Berridge, 1993). Indeed pharmacological investigation of striatal areas support a distinction between wanting and liking (K. C. Berridge & Robinson, 2003). Recent experiments with dopamine deficient mice lead Berridge (2007) to argue that the putative dopaminergic reward signals instead signal degree of desire (i.e., wanting or, in their parlance, incentive salience). Tyrosine hydroxylase knock-out (DD) mice, mice who lack dopamine, still can learn reward contingent tasks, but they do not act on that (latent) learning until dopamine is restored (K. Berridge,

2007). In a similar vein, DD mice also display a reward preference when given the choice of a sweetened water versus untreated water, but unless dopamine is restored their overall desire for both is greatly decreased. Based on these findings they argue that dopamine is not necessary nor sufficient for reward driven learning. By combining studies of addicts (and their non-human animal equivalents) who display increased "wanting" of drugs but not "liking" (people rate the experience as no more pleasurable than controls, mice consume no more of the substance that controls) K. Berridge (2007) argue that phasic dopamine signals a stimulus' incentive salience. Conclusions based on DD mice, though, should be cautious. DD mice are extraordinarily lethargic. To pep them up, caffeine is administered. Caffeine, through a cascade driven by adenosine A2A and cannabinoid CB1 receptors in ventral striatum, can have biochemical effects similar to dopamine (Lazarus et al., 2011; Rossi et al., 2010).

Still, neglecting concerns about caffeine, the experiments in DD mice are quite damning to the theory that dopamine plays a casual role in stimulus-response learning. There is, however, a substantial body of work supporting causation. Administering human and non-human animals dopamine antagonists adversely affects stimulus-response learning (Pizzagalli et al., n.d.), as does lesioning either the VTA/SNc or portions of the striatum. Complete lesions of the striatum prevent stimulus-response learning (Packard & Knowlton, 2002). This is relevant as it is the interaction between phasic dopamine and the striatum that is proposed to guide (drive) stimulus-response learning. Administering dopamine agonists, or the readily converted dopamine precursor L-DOPA, leads to increased Pavlovian instrumental transfer as well as re-

sponse vigor (Winterbauer & Balleine, 2007), both of which are thought to be facilitated by the interaction between phasic dopamine and activity in the ventral striatum. Parkinson's patients when off medication, and so lacking striatal dopamine, show marked decreases in stimulus-response learning (Pizzagalli et al., n.d.). These same off-medication patients show an enhanced (compared to normal age and intellect matched controls) capability to learn from negative feedback, suggesting their ability and desire to act is intact (Frank et al., 2004).

While it not clear how to reconcile, theoretically or functionally, the evidence for incentive salience and reward prediction, it might not be necessary. K. S. Smith et al. (2011) showed distinct semi-overlapping tuning in VTA/SNc for both reward prediction, incentive salience as well as with measures of the animals enjoyment of the in-task reward (i.e., liking). There may in fact be no one correct theoretical accounting; the neurons of the VTA/SNc may signal instead a family of functions - for other supporting examples see, Ito and Doya (2011); K. S. Smith et al. (2011); Bornstein and Daw (2011); Bromberg-Martin, Matsumoto, and Nakahara (2010); Matsumoto and Hikosaka (2009).

Thinking About Thinking Rewarding Thoughts

As was stated at the outset, cognition alone can generate activity similar in appearance and effect to that seen following primary and secondary rewards. For example, Tricomi and Fiez (2008) showed ventral striatum BOLD signal changes in a declarative memory task in which subjects were initially trained with feedback to distinguish 60 correct from incorrect word pairs. In the subsequent two rounds

explicit feedback was withheld but activity in the caudate was observed when correct pairings were matched based on memory alone, i.e goal achievement led to strong activity. In two economic decision making tasks strong ventral striatum signals were observed when participants were required merely to imagine or consider alternative outcomes (Hayden et al., 2009; Lohrenz et al., 2007). Information about the future is rewarding as well; Bromberg-Martin and Hikosaka (2009) showed that complex visual clues about an upcoming outcome were sufficient to cause bursts of firing in the VTA/SNc. Inversely, neutral stimuli can prevent decreases in responding that normally accompany repeated delays in reward presentation (Reed, 1992), a phenomenon that was reproduced using a robotic rat, wherein the neutral stimuli were treated as intrinsically rewarding (Fiore et al., 2008).

Informative, or to change terms to keep with other literatures, salient[6], stimuli have been observed to have rewarding-like effects in people as well, though supporting data is limited to fMRI experiments. Striatal BOLD increases have been observed in response to infrequently presented flashing images (Zink, Pagnoni, Martin, Dhamala, & Berns, 2003), and unexpected alarming tones (e.g., a siren replacing a constant 60Hz tone) (Zink, Pagnoni, Chappelow, Martin-Skurski, & Berns, 2006). Activity in the ventral striatum appeared in these tasks due to the stimuli alone, while dorsal activity was seen only when the stimuli had behavioral relevance. A control task suggested this increase was not in response to the additional motor demands of the response but was, the authors argued, due to the increased salience of the active versus passive condition. The context of behavioral responses made the reward more

[6]Not to be confused with the "incentive salience", discussed above.

salient. A similar dorsal/ventral division has been reported when comparing passive reward receipt to reward receipt requiring a response (O'Doherty et al., 2006) though these were attributed directly to the need for an instrumental response and not (necessarily) contextual salience. Additionally, like reward expectations, salience-related activity scaled with intensity (Zink et al., 2006).

Novel (but not necessarily salient) stimuli also elicit reward prediction-like dopaminergic firing in monkeys (Blatter & Schultz, 2006) and in people (Bunzeck & Düzel, 2006). Indeed, reward and novelty appear interchangeable. Guitart-Masip, Bunzeck, Stephan, Dolan, and Duzel (2010a) showed that when novel images precede rewarding outcomes ventral striatal activity is enhanced compared to reward alone. Rewards preceding a visual stimulus or word-pair leads to enhanced memory for that pair (Lisman & Grace, 2005). Building on that finding Wittmann, Bunzeck, Dolan, and Düzel (2007) showed enhanced recognition of natural scenes when images were preceded by novel images. This effect was reproduced using high-resolution imaging of the VTA/SNc, with that area demonstrating a marked reward prediction signal during the task (Krebs, Heipertz, Schuetze, & Duzel, 2011). This effect was extended further to the anticipation of novel images, similar to reward anticipation studies of striatal function (Knutson, Adams, Fong, & Hommer, 2001). Novelty driven exploitation/exploration decision making relies on the striatum as well (Wittmann, Daw, Seymour, & Dolan, 2008).

In sum, task completion, imagined rewards, neutrally valued informative cues, behaviorally salient but non-rewarding events, as well as novel images, have all been shown to act as reinforcers, and to stimulate the dopaminergic midbrain into phasic

firing. None of these, individually or as a group, can be explained parsimoniously as secondary rewards[7]. They were never *statistically regularly* paired with a primary reward.

Generally Generalizable.

In a recent article Wimmer, Daw, and Shohamy (2012) made the argument that a stimulus' reinforcement derived value must be generalizable to other similar stimuli as there are far too many possible states in the world to explore each individually, *and* therefore you rarely if ever encounter the same exact same stimulus twice. In essence, they make an argument that categories are necessary for reinforcement learning to be a reliable guide in our complex changing world. That reasoning is now extended. There are at least as many outcomes as there are stimuli, i.e., any stimulus can be an outcome, in principle anyway. There are therefore too many possible outcomes to ever experience them all, *and* you'll rarely if ever see the same outcome twice. As a result, you often can't predict exactly what will happen. Therefore like stimuli, outcome expectations must generalize. As desirable outcomes are rewarding (see p19 in the *Introduction*), reward representations generalize. Skipping ahead a little, it is hypothesized that the act of generalization will impact reward prediction error calculations.

Needed similarities. And indeed many of the cognitive rewards outlined above may employ or even require generalization, often in the form of similarity assessment.

[7]Nor as primary rewards, but this is a definitional problem.

For example, to find that successful task completion is rewarding, Tricomi and Fiez (2008) used a cued recall task, which in part relies on familiarity (Jacoby, 1991) and thus similarity (R. Nosofsky, 1988). Likewise studies of fictive rewards – rewards created only in the participants imaginations – require an act of inference; past desires must be extrapolated into a present experience (Hayden et al., 2009; Lohrenz et al., 2007). Roesch et al. (2007) demonstrated rewarding inferences in rats as well. When novel stimuli were presented in the context of a familiar task with the same action options previously available, the dopamine spikes that resulted were significantly correlated with the most valued past action. Likewise assessing both salience and incentive salience requires participants to extrapolate past context and goals and into future experiences. In fact Zink, Pagnoni, Martin-Skurski, Chappelow, and Berns (2004), showed that rewards sans salience elicited no striatal BOLD activity. This finding suggests that rewarding activity must be driven by more than just the current hedonic stimulation.

Birds do it. The argument for cognitive-type rewards and their hypothesized category representation is novel, however there are several non-human animal studies examining the generalizable properties of secondary reinforcers, though the neural mechanisms for such are unexamined. As an example, Guttman (1956), varied an instrumentally conditioned 570 nm light from 480-610, showing that while the bird's pecking rate (i.e., response vigor) decreased as one moved farther from 540, the birds still responded. They generalized across the wavelengths. When novel variations of conditioned stimuli from two sensory modalities were mixed similar graded changes in vigor were observed (Guttman, 1956). Pigeons' capability for gener-

alization are not limited even to mixtures of simple cues: (Nakamura, Ito, Croft, & Westbrook, 2006) taught pigeons to discriminate categories of male and female birds and J. D. Smith et al. (2011), taught pigeons perceptually complex yet abstract categories. Interestingly though, not all experiments were effective at producing generalization behavior. Some combinations produced no responses at all (Blough, 2001; Simmons, 2008; Urcuioli, 2001). Interestingly, the degree of generalization was not uniform across stimuli, e.g., generalization between shapes decreased more rapidly than colors (Shepard, 1987). The source of this variability remained unknown until Shepard's seminal insight.

Curves and categories. Shepard (1987) demonstrated that the variability among instrumental response generalization curves in pigeons and other animals could be accounted for if one no longer measured generalized responding by an objective metric (e.g., wavelength) but instead estimated the psychological space of the animal's perception. In his own words: "I propose to start with the generalization data and to ask: Is there a unique monotone function whose inverse will transform those data into numbers interpretable as distances in some appropriate metric space". The metric space he proposed required "[...] a very strong additivity condition: For each subset of three points, the distance between the two most widely separated points equals the sum of the distances between those two points and the third point that lies between them". After making some additional assumptions[8] Shepard shows that the large variability among the many objective spaces

[8] 1. that a given stimulus is equiprobable over the psychological space and 2. that categories in the space are centrally symmetric and convex

he examined only minimally impacted the estimated (negative exponential or Gaussian) psychological space. In other words, in many cases the objective space can be mapped to psychological distances by simply taking the negative exponential of their Euclidean distances[9]. Shepard's work suggests that whether the psychological space is exponential or Gaussian depends on whether the stimuli were distinct or perceptually confusable, which was partly supported in theoretical studies of perceptual noise on generalization (Ennis, 1988). Outside of Shepard's theory, though, the Gaussian metric has substantial empirical and theoretical support in studies of human category learning, even when confusability of the stimuli is low (R. Nosofsky, 1985; Medin & Schaffer, n.d.). As such the divide between exponential (as seen in the Pigeon studies above) and Gaussian metrics (often seen in human studies) remains unbridged.

Based partly on Shepard's work the fields of perceptual and cognitive categorization have since made an ever evolving study of how animals create and use categories. However there are still substantial controversies with no one consensus view. The two extreme theoretical positions are exemplar models and prototype models (Ashby & Maddox, 2005). Exemplar models, in their simplest form, are feature counters, i.e., their categories are completely defined as the sum of previously observed information (R. M. Nosofsky, 1988). Prototype models instead hold only information about one ideal template (Rosch, 1973), often implemented as probability density (Ashby & Alfonso-Reese, 1995), i.e., a set of sufficient statistics over the relevant information. A common example is the mean and covariance, assuming prototype

[9]Though this geometrical relation is not necessary. Shepard's insight can be re-derived using information theory (Chatera & Vitányib, 2003)

following matching a Gaussian distribution. In between the two extremes are hybrid approaches, of which there are many. These include using mixture models (Rossee, 2002), neural networks implementing self-organised maps (Love, 2004), and other clustering techniques (Kruschke, n.d.) as well as kernel methods (Jakel, Scholkopf, & Wichmann, 2008), among other increasingly esoteric approaches (Martin, Griffiths, & Sanborn, 2012). There are many more examples, as categorization has been an active area of investigation for over 40 years.

This deep mathematical literature is further complicated by empirical investigations of human category learning, which suggest that regardless of the underlying representation, humans treat different category structures in different ways (Ashby & Maddox, 2011). In their hands, category learning has been subdivided into systems to handle rule-based (i.e., verbalizable), information-integration (implicit, forces consideration of > 1 dimension of information), prototype-distortion (perceptual categorization), and unstructured category (categories without perceptual or semantic overlap, akin to sets in that their only necessary commonality is the fact they're grouped together). Like finding the neural correlates of these systems (Ashby & OBrien, 2005; Ashby & Ennis, 2006), integrating them with the many formal approaches is in its early days (Ashby & Maddox, 2011).

Why Are We Here Again?

Are cognitive rewards represented as categories in the human brain? And does such a representation impact the reinforcement learning process? To start to answer these two questions, participants completed a stimulus-response task where

classical rewards (e.g., "Correct!" or "Win $1.") were replaced with pre-trained perceptual categories, one category for gains and one for losses. The reward for each trial was a unique, never before or again experienced, examplar from one of the two reward categories, distinguishing this task from higher-order conditioning paradigms were the same stimulus is repeatedly paired or presented. That is, as each reward was a new abstract shape (a black and white sinusoidal grating), its value had to be in some way inferred if the subjects were to learn the correct stimulus-response relations; they learned these relations quickly. So then, that some inference occurred is logically necessary, however the models examined here postulate a mechanism for that inference – rewards as categories.

Behavioral and fMRI data collected under this task was then compared to a series of computational and functional connectivity models representing various possible algorithmic strategies for reward inference and assessing category membership. These analyses are spread across two chapters (i.e., Chapters 2 and 3 below). The first covers the details of the task itself, the computational models and their parameter optimization. The second covers the fMRI methods and basic brain-mapping analyses, as well as the BOLD-signal-to-model comparisons. The final chapter concludes that rewards are categories, as demonstrated by behavioral, fMRI, and modeling results.

Rewards as categories represents a substantial change in how rewards are conceived, and modeled: the primary, to secondary, to higher-order conditioning paradigm is incomplete, perhaps even incorrect. Secondly, these experiments represent the first confirmation that the complex phasic activities reported at the neuronal

level in the VTA/SNc (H. Kim et al., 2006; Matsumoto & Hikosaka, 2009; J. D. Smith et al., 2011) contribute meaningfully to the BOLD signal, which represents the aggregate activities of thousands and thousands of cells – a notable advance that suggests analyzing detailed, multiple sub-population, neuronal models may be possible using fMRI.

Chapter 2 – Task and Models

This chapter has two parts, the overall aim of which is to show how a stimulus-response task was used to examine possible category representations of rewards, and then to described how learning in that task was modeled (using a set of reinforcement learning equations). To that end, and in that order, first the behavioral task is described and its results characterized. Second, the computational models are rigorously laid out, as are their results.

On Task

What they did, and when. The behavioral task each participant completed consisted of two parts. Depicted in Figure 1. (*top*), the first was a passive learning task wherein participants learned two rewarding perceptual categories by viewing randomly selected black and white sinusoidal gratings. Each grating (which was on-screen for 2 seconds) was followed by "Gain $1" or "Lose $1" in, respectively, green or red letters (1 second). The width of the grating's lines and their angles was derived from an information integration (category) distribution (Figure 2; borrowed from Spiering and Ashby (2008)). The disappearance of each grating and appearance of the reward was separated by an empty grey screen (1 second). Each trial terminated in a fixation cross (lasting at least 0.5 seconds). In total then, each trial lasted a total of 4.5 seconds. The trials for part 1 were spread over an initial training period completed outside the scanner, lasting 126 trials, and an in-scanner refresher lasting 45 trials. Prior to beginning training participants were, after some preliminaries,

instructed to, "Attend to the screen in order to learn which types of gratings indicate wins and which types indicate losses". To minimize any stimulus specific effects, the category parameter distribution (Figure 2) to reward (i.e., "Gain $1" or "Lose $1") mapping was randomized for each participant.

Part 2 was a stimulus-response task that replaced classical rewards with an appropriate grating from task 1 (Figure 1, *bottom*). Gratings matching the Gain category were used for positive reinforcement, while gratings indicative of losses were used as negative reinforcers. Each trial began with an abstract black and white "tree" stimuli (left most image in bottom of Figure 1). Each "tree" deterministically belonged to one of two arbitrary response categories ("q" or "w"). Subjects indicated their response by button press using either the right or left index fingers on a magnet-compatible response box. The response window lasted up to 2.5 seconds, but ended as soon as a response was made. Immediately following response the "tree" was replaced with a blank grey screen, which was on-screen for half a second and was replaced with a feedback screen. If the response was correct a *never before experienced* exemplar grating from the gain distribution was used; if the participant was incorrect, a new loss grating appeared instead. The use of novel gratings forced the subjects to classify each grating prior to inferring its value. This necessary inference made these rewards incompatible with primary or secondary definitions. If no response was made, or the wrong button was pressed, the subject's reward was replaced with, "No response detected" (these trials were excluded from further analysis). Feedback always lasted for 1 second and was terminated by a fixation cross (0.5 seconds).

For the instructions in part 2 participants were told, "Each tree belongs to

either category q or w. Which is the correct answer though is random. The shape of trees is meaningless. To learn the correct response for each tree you must start by guessing. Use what you learned about the rewarding properties of the gratings from part 1 to learn the right responses. Remember, a random subset of the Gains and Losses are real. These mostly determine how much you'll earn for your participation. So try and earn as much money as possible". Instruction for both parts were given orally by the experimenter using a script and Figure 1 as a visual aid.

Over the course of part 2, participants learned to classify 6 "trees", randomly selected at the start of the experiment out of a pool of 22 possible. Each of the 6 were experienced a total of 28-32 times for a total of 199 trials. The order of the trials in the second half of part 1 and all of part 2 was determined using a genetic algorithm designed to optimize fMRI signal detection, among other considerations. Most relevant to behavioral analysis, trials were in pseudo-random order with second order counterbalancing. For remaining details see p51.

As part of fMRI data acquisition, 18 participants completed both parts of the task (10 female, mean age of 24, ranging from 21 to 32). Participants were compensated at a base rate of $15.00 earning up to $30 more depending on behavioral performance. For the last 30 trials the participant was paid an additional dollar for every correct response and lost a dollar for every incorrect response. Before beginning experiment participants were told that some trials would count, but were not told till after which trials (i.e., the last 30). The highest payout was $45, i.e.,perfect performance, the lowest was 30, indicating near chance behavior for the last 30. The average payout was $40.23.

Of the 18 participants, two were removed all analyses as they demonstrated inverse learning (Figure 3, see *107* and *110*). Despite reporting a full understanding of the reward contingencies from part 1, in part 2 these participants displayed significant and consistent decreases in performance through time. Had this learning been in the usual direction it would have been considered better than average performance. In post task interviews both reported feeling as if they performed above average. Once they were informed of their inverse performance neither believed it. It seems possible then that both correctly learned the perceptual characteristics but mis-mapped the value labels, i.e., they got "Gain" and "Loss" mixed up. Post-experiment interviews further suggested that both the discarded subjects were under high personal stress. One subject, who was a PhD student, had his competency exams the next day. The other completed a 60 mile bike ride an hour prior to participation. Combined these participants data suggest that the perceptual and verbal characteristics of the reward categories are independently accessible, and that the verbal label may be more labile than the perceptual distributions. However given the other participants consistent positive performance and rapid learning it seems these two were a curious but isolated anomaly (Figure 3).

Well behaved results. On an individual basis the lower confidence interval[10] around the binomial fit of the accuracy data rose to above the chance level (0.5) by the last trial, except for participant 103, who did not learn (Figure 3). Many participants (11 of 16) greatly exceeded this minimum criterion, showing above chance

[10] All confidence interavals were bootstrapped estimates calculated suing the Hmisc package (`http://cran.r-project.org/web/packages/Hmisc/index.html`) in the R programming langauge (v2.15.1; `http://www.r-project.org/`)

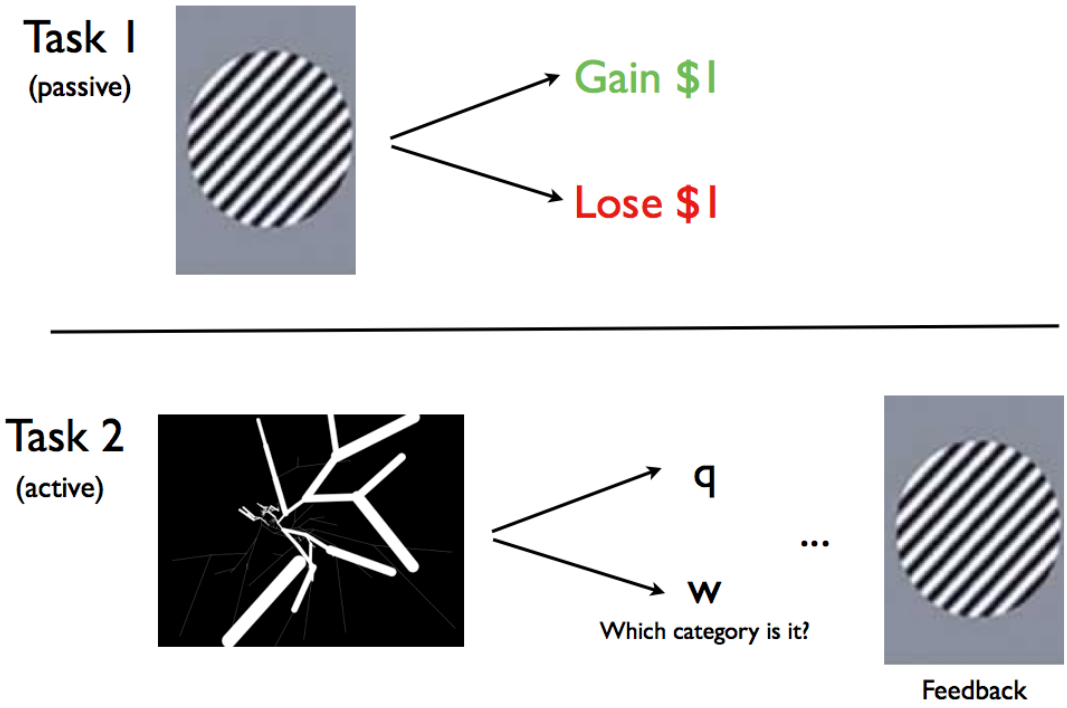

Figure 1. Depiction of the behavioral task. The *top* depicts part 1, the passive learning of the reward categories. The *bottom* depicts part 2, the stimulus-response learning phase.

learning by trial 10, and nearly all (14 of 16) exceeded chance by trial 20 (Figure 3). These individually good performances are reflected in the participants' aggregate performance, which was well above chance by trial 5 (Figure 5). This aggregate learning rate is consistent with past work in the lab using verbal or monetary feedback (Seger, Peterson, Cincotta, Lopez-Paniagua, & Anderson, 2010; Seger & Cincotta, 2005), as it is with other results in other labratories (O'Doherty et al., 2003; Ramnani, Toni, Josephs, Ashburner, & Passingham, 2000; Aron et al., 2004).

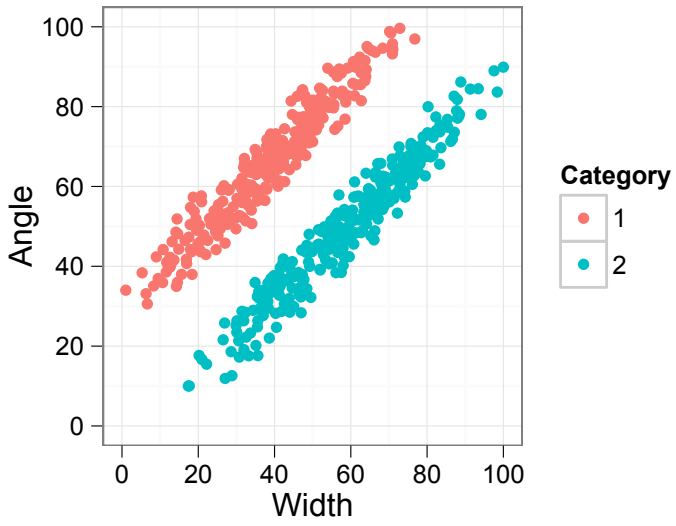

Figure 2. The two sinusoidal grating distributions for the information integration (II) category distributions. As II categories span the diagonal of the gratings parameter space (line width and angle successful learning requires consideration of both dimensions preventing participants from solving the categorization problem with simple rule-based strategies.

(E. E. Smith & Grossman, 2008; Poldrack & Foerde, 2008; Ashby & Ennis, 2006; Ashby & Maddox, 2005; Ashby & OBrien, 2005)), indicating that the rewarding categories in present study are behaviorally similar to classical rewards. The consistency between classical rewards and this task were also reflected in the reaction time measures, which showed a 200 ms decrease over time, bottoming out near 850 (for individual averages see Figure 4 and for overall performance see Figure 6). Responses in similar, classically rewarding tasks end with reaction times near 700-800 milliseconds and show similar rates of decline. The 50-100 ms possible difference in reaction times may be due to the increased difficultly of classifying the rewards compared to simply reading the value of the outcome (e.g., “Gain $1”).

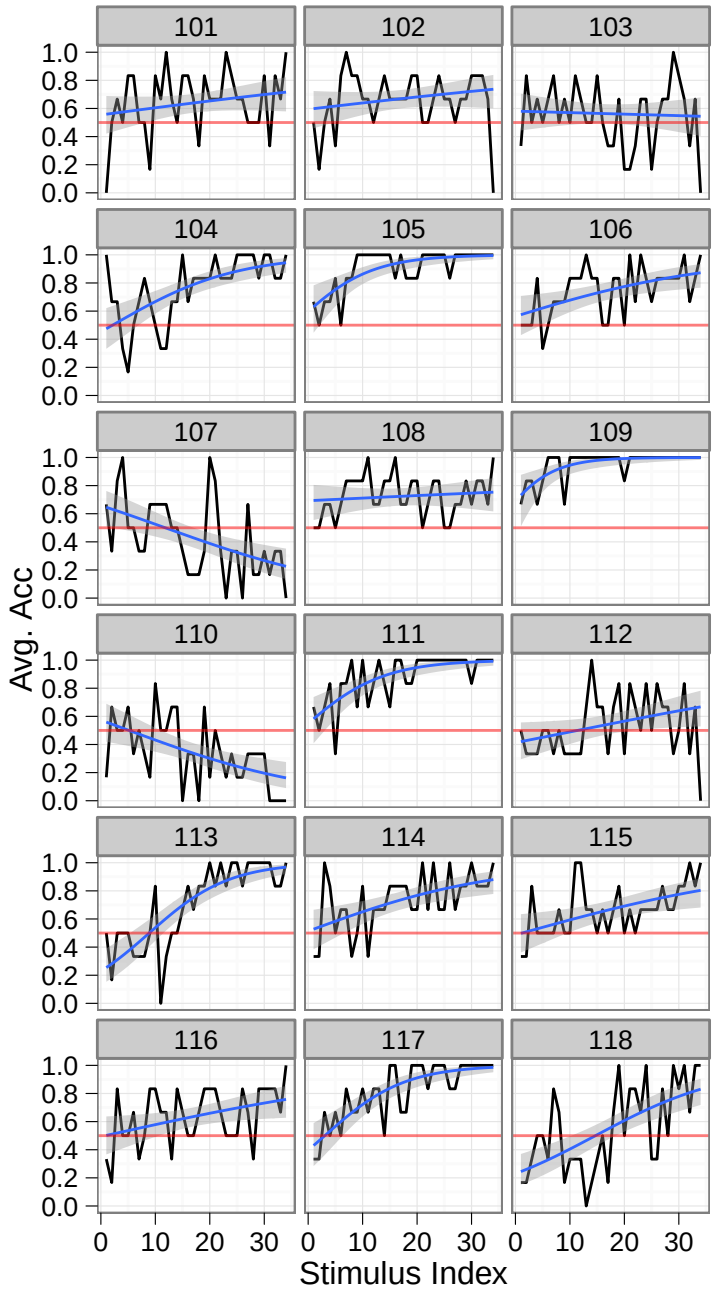

Figure 3. Average accuracy for each participant (black), averaged for all 6 stimuli by trial (i.e., Stimulus Index), blue line and the grey area represent a binomial regression fit of the data and bootstrapped 95% confidence intervals, respectively.

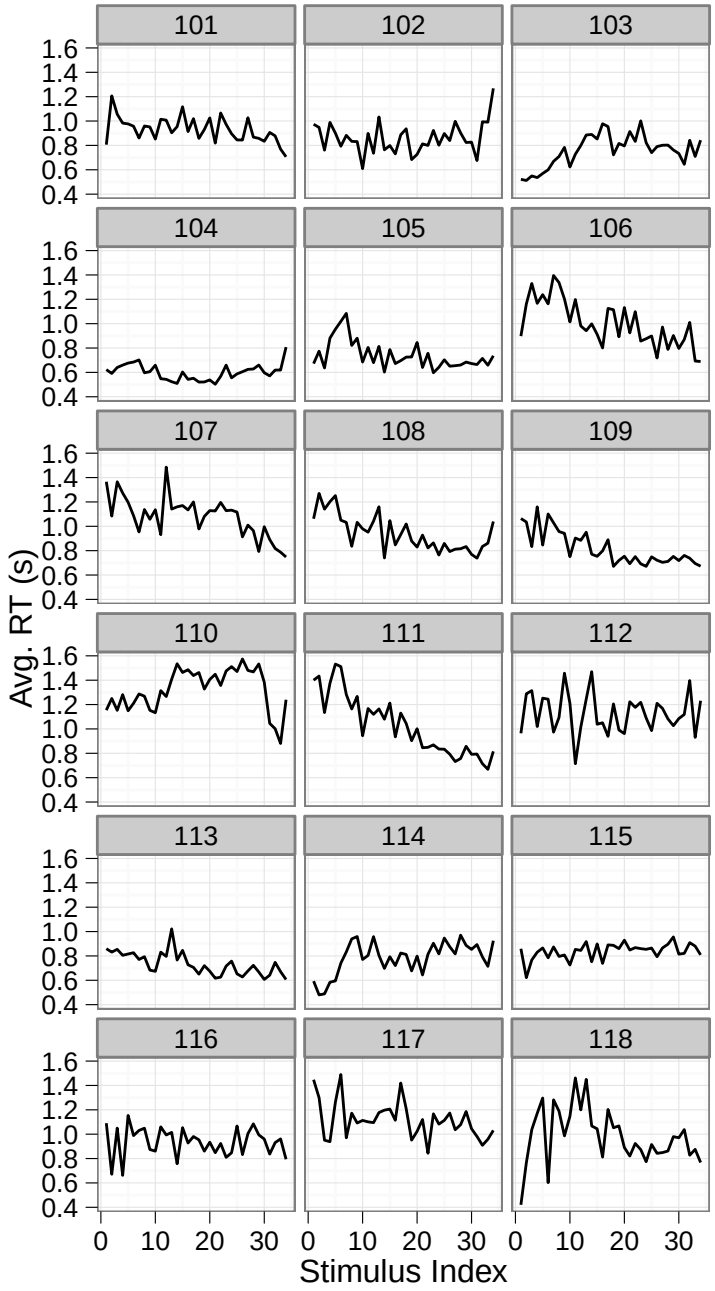

Figure 4. Mean reaction time for each participant, averaged for all 6 stimuli by trial (i.e., Stimulus Index).

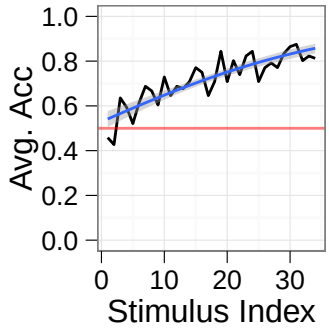

Figure 5. Mean accuracy (black), averaged over all participants and stimuli, plotted by trial. The blue line and grey area represent a binomial regression fit of the data and bootstrapped 95% confidence intervals, respectively.

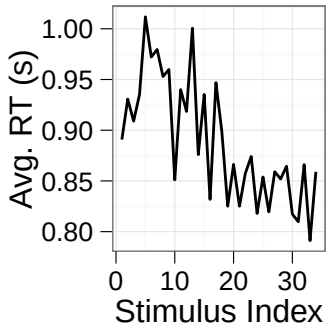

Figure 6. Mean reaction time (black), averaged over all participants and stimuli, plotted by trial. The blue line and grey area represent a binomial regression fit of the data and bootstrapped 95% confidence intervals, respectively.

3 Models and 2 Codes

Our three models. Three Rescorla-Wagner-like models were constructed, each using a distinct reward representation. The first model treated rewards identically to a classical reward (e.g., a 0 for a loss or 1 for a gain). The second and third devalued each reward based on how similar it was to the category mean. This similarity metric is (necessarily) simplistic. As was reviewed on p24, there are many proposed models for how the categories are represented. Likewise, our understanding of the computational implementation of the category learning systems is only just underway (Ashby & OBrien, 2005; Ashby & Ennis, 2006). As such there is no obvious way to interlace the hypothesis of rewarding categories with the category learning systems or with the many models of categorization they rely on. Avoiding such ambiguity, a simpler route driven by Shepard's basic finding (Shepard, 1987) was taken. Whatever the representation and/or category learning system that (may) implement rewarding categories, Shepard's work insists they show an exponential or Gaussian decline with similarity. For simple stimuli like a light or tone, similarity is measured from the initial training prototype (Guttman, 1956). However as the task does not have a singular prototype the mean of the parameters for all training trials (i.e., part 1) was used in its place. This simple substitution makes categories identical to the simplest of the prototype category representations (Rosch, 1973; Ashby & Alfonso-Reese, 1995), making this a parsimonious yet literature driven first attempt to quantify similarity of rewards. Therefore, for model two similarity decreased exponentially measured from the training mean. While for model three it decreased along a normal Gaussian (for complete mathematical detail see p39).

Codes and fits. For each participant and model, the two free parameters (α, which controls each model's rate of learning and β, which controls the steepness of the action selection criterion) were fit using an exhaustive[11] maximum log-likelihood search. Additionally each model was run using two separate reward coding schemes. In the first scheme Gains were valued as 1 and losses as 0. The second, which was based on the bivalent monetary value of the rewards, uses 1 and -1 for gains and losses respectively. The first scheme is universally used in human and animal modeling studies as well as in machine learning, and was Sutton and Barto (1998) recommended scheme. However recent recordings of dopaminergic neurons in monkey suggest that the true reward codes are complex, even perhaps redundant (H. Kim et al., 2006; Matsumoto & Hikosaka, 2009). Among other schemes, they reported firing consistent with a bivalent reward code. Using the second scheme is merely a first, albeit simple, step in incorporating the potential complexity of dopaminergic firing as observable by the fMRI BOLD signal and in human subjects.

The incantations. To restate more formally, a Rescorla-Wagner-like model's value updates were defined by,

$$V(s_t, a_t) \leftarrow V(s_t, a_t) + \alpha\delta \tag{1}$$

$$\delta = r_{(classic,t)} - V(s_t, a_t) \tag{2}$$

where δ is the reward prediction error, s is a stimulus, or state (or which there were 6), and a is an action (either "q" or "w") and $r_{classic}$ (the numerical representation

[11] With a 0.05 precision, ranging from 0-1 for α and 0-5 for β

of the rewards) can be coded as either

$$r_{classic} = \{1, 0\} \tag{3}$$

or

$$r_{classic} = \{1, -1\} \tag{4}$$

but where $r_{classic}(t)$ may also be replaced with

$$r_{exp} = r_{classic} S_{exp} \tag{5}$$

or

$$r_{gauss} = r_{classic} S_{gauss} \tag{6}$$

where D is the Euclidean distance from the width (w) and angle (θ) of that trials reward category to the average values from the pre-training ($\bar{\theta}$ and $\bar{W}$; see discussion on part 1 on p29)

$$D = \sqrt{(\bar{\theta} - \theta)^2 + (\bar{W} - w)^2} \tag{7}$$

is transformed to a Shepard-like similarity metric (Shepard, 1987):

$$S_{exp} = e^{-D} \tag{8}$$

$$S_{gauss} = e^{-D^2} \tag{9}$$

Consistent with past work, all values are initialized at 0 (Beierholm, Anen, Quartz, & Bossaerts, 2011; Bischoff-Grethe et al., 2009; Gershman, Pesaran, & Daw, 2009)

$$V_{initial} = 0. \tag{10}$$

and values are transformed to response selection probabilities via the softmax distribution (Sutton & Barto, 1998; O'Doherty et al., 2003).

$$p_1(s_t, a_t) = \frac{e^{\beta V_1(s_t,a_t)}}{e^{\beta V_1(s_t,a_t)} + e^{\beta V_2(s_t,a_t)}} \tag{11}$$

where V_1 and V_2 are the values for the two response options (i.e., "q" and "w"). During maximum likelihood parameter selection (p39), the $p_1(s_t, a_t)$ values from each trial and some test parameters (α_{test} and β_{test}) calculate the log-likelihood by,

$$L = \sum log_e(p_1(s_t, a_t)) \tag{12}$$

Fits and plots. On average none of the three models fit the accuracy data better than the rest (Figure 7). For brevity's sake each model will be referred to as "none", "exp" and "gauss", corresponding to Eq 2, Eq.5, and Eq. 6 respectively. Nor did the coding scheme impact the fits ("acc" and "gl", matching respectively Eq. 3 and Eq. 4; Figure 7)). For "acc" the step size parameter ("alpha" in Figure 8 matching α in Eq. 1) increased in "exp" compared to the other models. This increase was expected as the exponential similarity metric can sharply decrease the magnitude of each value update, requiring an increase in α to compensate. A similar trend was

observed for the temperature parameter ("beta" in Figure 9, matching β in Eq. 11). The more equiprobable each action is the larger the temperature parameter. As such the increase for "exp" means participant's choices are more likely to change from trial to trial, which is again consistent with a decrease in update magnitudes.

Importantly the intra-subject variability in α and β was low, as demonstrated by the small standard error of both parameters (Figure 8 and 9). Consistent parameter estimates between subjects support the use of subject-level parameters in the fMRI analyses, which in other hands have been reported to be too noisy to be reliable (Daw et al., 2011; Seymour, Daw, Dayan, Singer, & Dolan, 2007; O'Doherty et al., 2003). Using subject-level parameters is a step crucial in assessing and maximizing model quality. The goal of any model of human behavior is to make good predictions for individual cases not just for aggregates of tens or hundreds of participants (Daw & Courville, 2007). However aggregates prediction is the norm in reinforcement learning models of human behavior (for examples see, Daw et al. (2011); Seymour et al. (2007); O'Doherty et al. (2003)).

The fit reinforcement learning models for every participant and coding scheme can be found in Figure 10 - 13. In the traditional formulation, where similarity does not impact reward value (e.g., see Figure 10 or 11) the reward prediction error decreases with learning, eventually plateauing at 0, for strong examples see subjects *102*, *105* and *111* in the "none" column of Figure 10. In contrast, both similarity models, "exp" and "gauss", never fully plateaued (again see Figure 10). In the context of the models, this is expected. Each grating will, in all likelihood, be non-identical to the mean. However as the parameter mean is the asymptotic expectation,

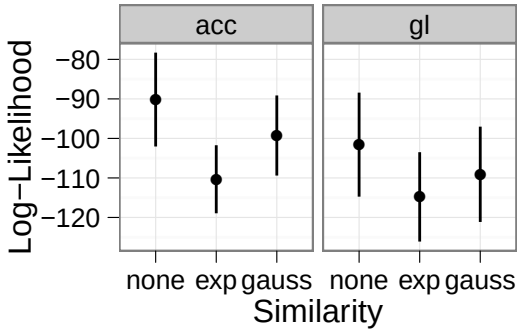

Figure 7. Average negative log-likelihoods for each of the models and coding schemes. Error bars represent standard errors.

reward prediction errors happen even after learning is complete. In the big picture, this is desired model behavior. The similarity adjusted model's are trying to capture the case where a reward's value may vary in ways that are not predictable *a priori* given the massive multiplicity of possible outcomes it is unlikely to find the same one multiple times. Small prediction error should then continue without end, as happens in these models.

When comparing "exp" and "gauss" models, you'll see the former appears to have lower magnitudes (Figure 10). Examining density plots composed of all participants data confirms this observation (Figure 14). The density plot also reveals that "exp" prediction errors are diminished more rapidly than their "gauss" counterparts (a pattern most clearly scene in the left panel of Figure 14).

Regardless of model class, between the two reward codes there are substantial differences in reward prediction behavior. The $\{1, -1\}$ (denoted in these plots as "gl") scheme leads to substantially more negative deflections that the $\{1, 0\}$ scheme

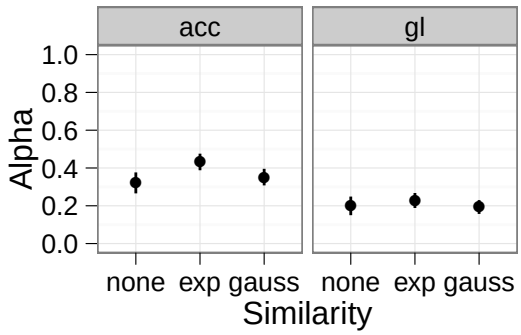

Figure 8. Average alpha values for each of the models and coding schemes. Error bars represent standard errors.

("acc") (compare Figure 11 and 10 as well as the *left* and *right* panels of 14).

Value estimates for the two similarity adjusted rewards (i.e., "exp" and "gauss") were generally less than the alternative classic model ("none" in Figure 12 and 13). However, as consequence of their reduced dynamic range, the similarity adjusted model's value terms grew more rapidly, for example examine participants *105* and *109* in Figure 12. In these cases both the similarity value terms approached their maximum by trial 50 whereas "none" (the unadjusted term) took until trial 150. That is, taking into account the uncertainty of each reward's worth lead to an increase in the learning rate. This increase was independent of the reward coding scheme (i.e., see also Figure 12- 13).

The coding scheme's impact on the value terms was two fold. First, as losses lead to larger prediction errors using the "gl" scheme (-1 compared to 0), value increased faster for "acc" (see Figure 12 and 13 as well as 15). Second, and most strikingly, "gl" lead to negative value estimates of the undesirable choices (Figure 13

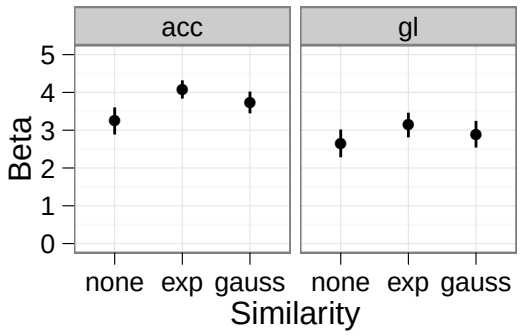

Figure 9. Average beta values for each of the models and coding schemes. Error bars represent standard errors.

compared to 12). While, so far as I'm aware, no reinforcement learning model of human or animal have considered negative value estimates, there is empirical support. As reviewed on p13, orbital frontal and ventral medial frontal cortices encode the absolute value of rewarding or punishing outcomes (O'Doherty et al., 2001; Hornak et al., 2004). As such, neural correlates of reinforcement learning derived negative value estimates might serve as an important link between theoretical and empirical findings on economic valuation. It might also serve as a link between reinforcement learning and affective/motivational processing (Knutson, Taylor, Kaufman, Peterson, & Glover, 2005; Delgado, Stenger, & Fiez, 2004).

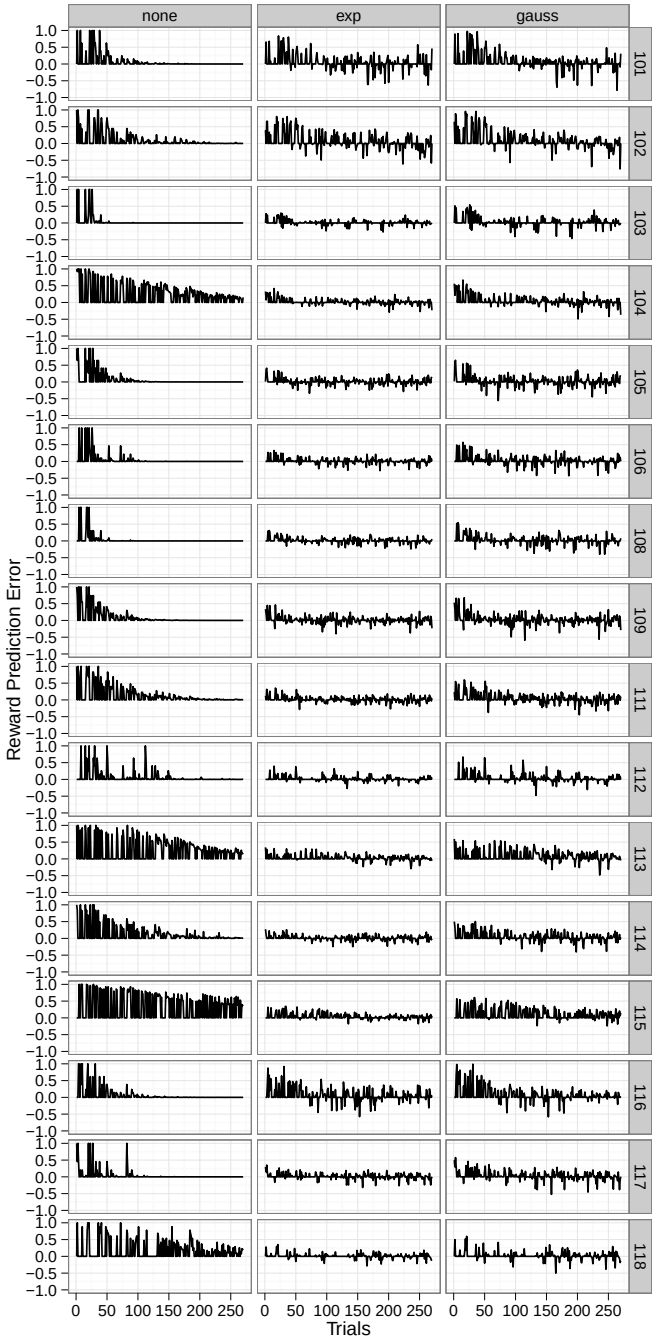

Figure 10. Reward prediction errors for each of the three models plotted for each trial in the experiment, based on the $\{1, 0\}$ coding scheme, which also represents the min-max range of the y-axis. Each row is a single subject's data. Each column matches one of the three models, classified by their similarity metric.

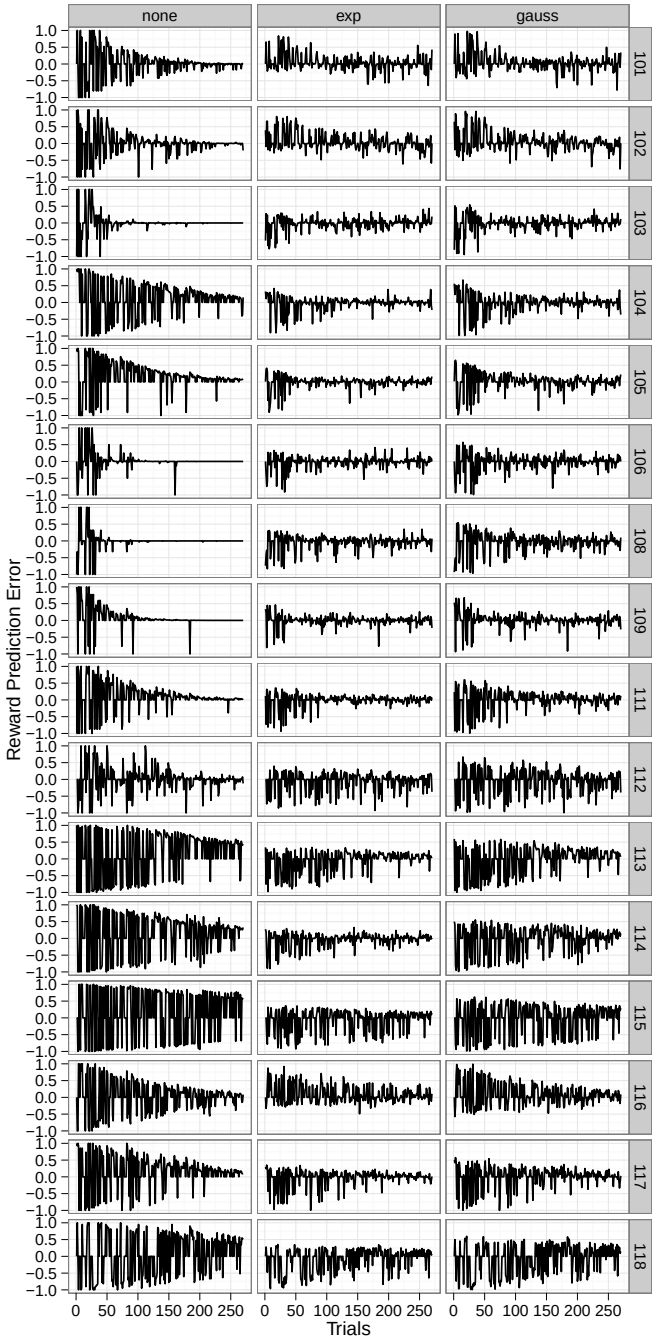

Figure 11. Reward prediction errors for each of the three models plotted for each trial in the experiment, based on the $\{1, -1\}$ coding scheme, which also represents the min-max range of the y-axis. Each row is a single subject's data. Each column matches one of the three models, classified by their similarity metric.

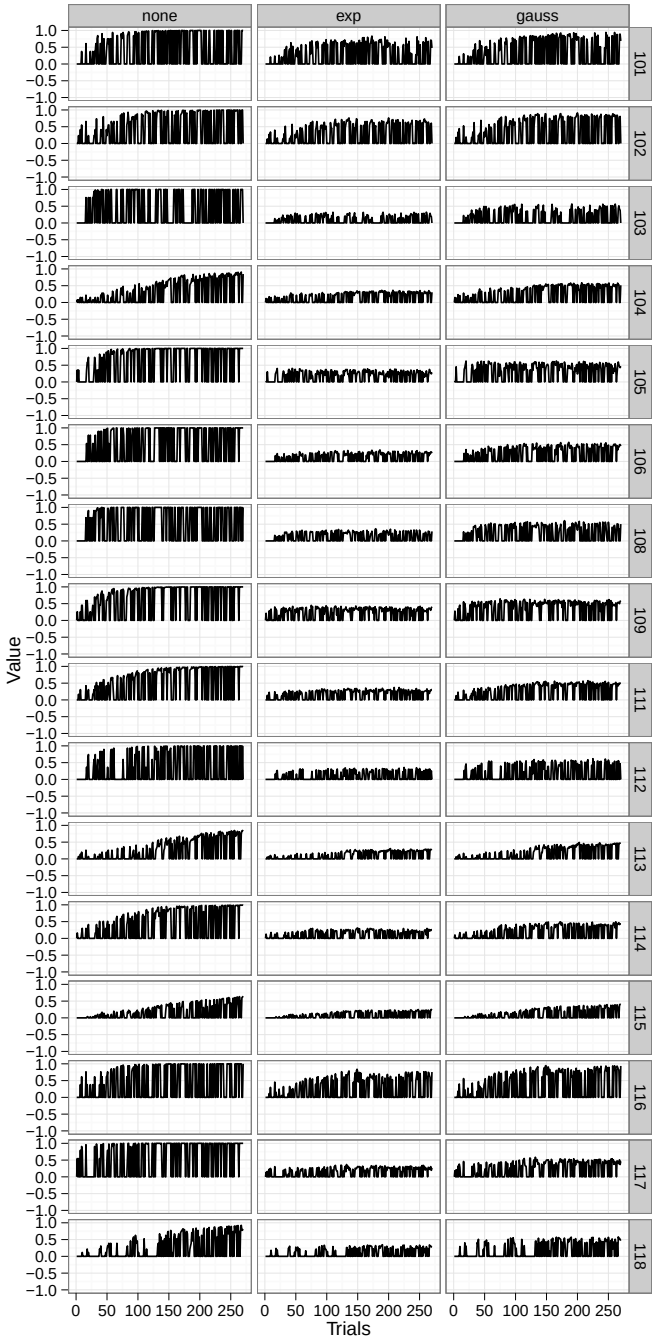

Figure 12. Value estimates for each of the three models plotted for each trial in the experiment, based on the $\{1, 0\}$ coding scheme, which also represents the min-max range of the y-axis. Each row is a single subject's data. Each column matches one of the three models, classified by their similarity metric.

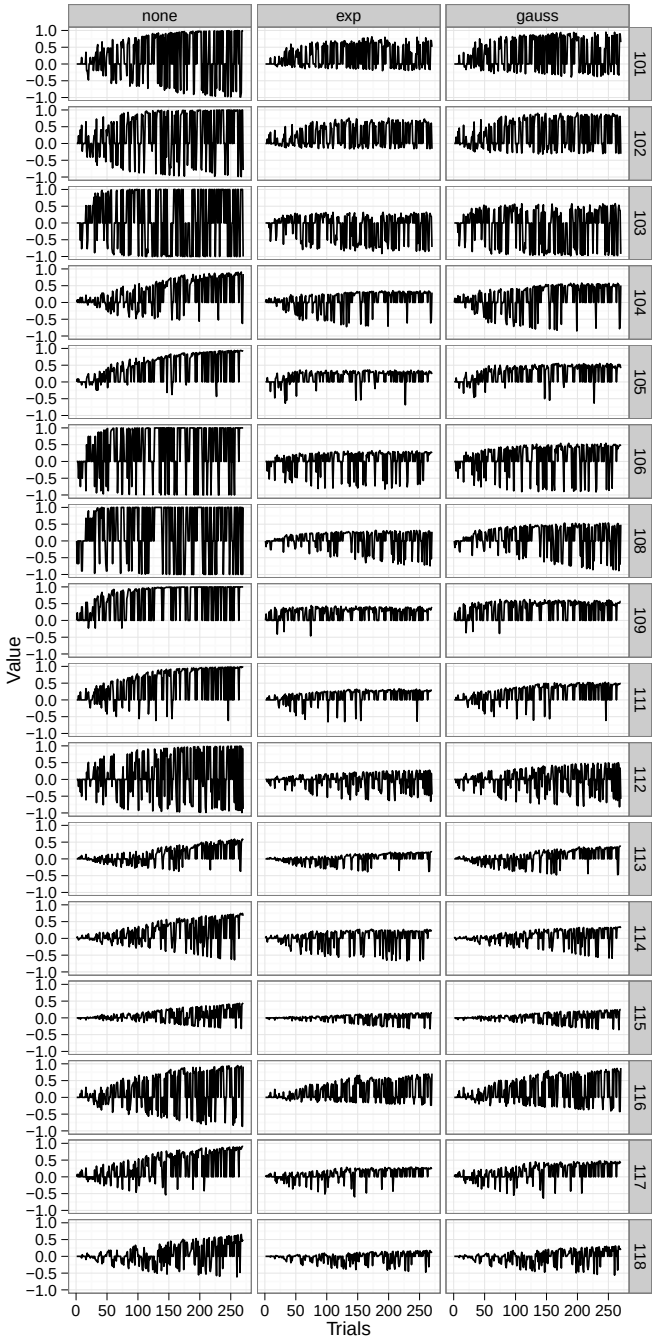

Figure 13. Value estimates for each of the three models plotted for each trial in the experiment, based on the $\{1, -1\}$ coding scheme, which also represents the min-max range of the y-axis. Each row is a single subject's data. Each column matches one of the three models, classified by their similarity metric.

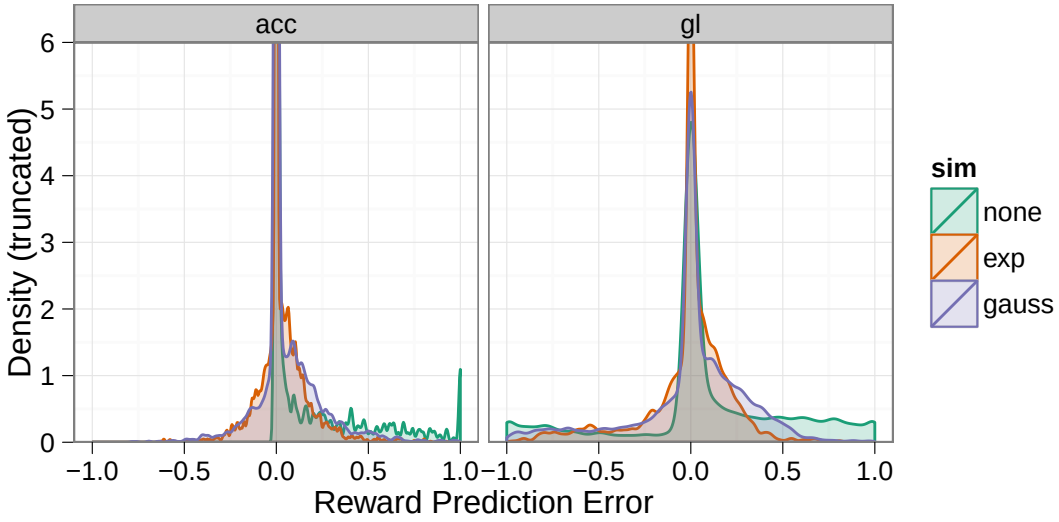

Figure 14. Density of reward prediction errors for all subjects. The y axis is truncated at 6 to allow clear visualization of non-zero values.

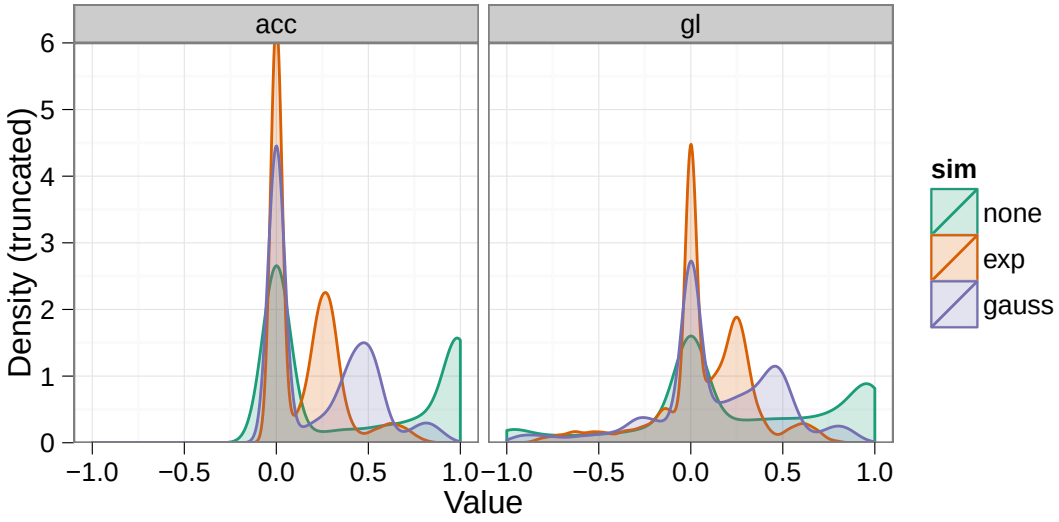

Figure 15. Density of value estimates for all subjects. The y axis is truncated at 6 to allow clear visualization of non-zero values.

Chapter 3 – fMRI analyses

This chapter has 4 parts, all address different aspects of fMRI data collection and analysis. First is the methodolical details of the fMRI functional (i.e., BOLD) and structural acquisition, as well as the preprocessing of that data. Little here exceeds or differs from current accepted fMRI data practices (Poldrack et al., 2008; Amaro & Barker, 2006; Bullmore et al., 1996). Second null-hypothesis test thresholded maps of BOLD activity are considered (i.e. whole-brain patterns of activity are discussed). The third section describes an alternative to the whole-brain analysis, focus on comparing BOLD time-courses from anatomical regions of interest to the reinforcement learning models from the previous chapter (p38). In this, the focus is on ranking models using information theory metrics, not trying to select *a* correct model as one might do in more traditional ROI analyses (for example, ? (?); Mars, Shea, Kolling, and Rushworth (2010)). Fourth is the results from the region of interest analyses described in part three. These results are, in general, highly supportive of rewards as categories, though the argument for that conclusion is held off until the next, final, chapter (p91).

An Acquisition

Data details. fMRI data was acquired at the Intermountain Neuroimaging Consortium (INC) facility located at the University of Colorado at Boulder on a Siemens Allegra 3T (whole body) scanner. All 18 right-handed participants were pre-screened for the typical fMRI exclusion factors (e.g., metal implants, mental

disorders, etc). High resolution anatomical data was acquired as a T1-weighted structural image, MPRAGE sequence, at 1x1x1 mm, (256 x 156 x 192) with a TR of 2530 ms, and TE of 1.64 ms, with a flip angle of 7°. All functional (i.e., BOLD) data was acquired with T2-weighted echo-planar imaging (EPI), at 2.29 x 2.29 x 4.00 mm (96 x 96 x 26), with a TR of 1500 ms, a TE a 25 ms, a flip angle of 75° and a FOV of 220 mm.

Four sets of functional data were acquired. The first was of the "refresher" for part 1 of the behavioral training (p29), spanning 241 volumes. The second and third spanned part 2 of the stimulus-responses learning task, divided into 2 (nearly) even sets lasting 390 and 394 volumes respectively (again see p29). The fourth scan featured repeated presentation of gratings from both reward categories, in a random order. The intent of this scan was to isolate rewarding activity outside the primary task. This localizer was not in the end useful (discussed on p55).

Preprocessed (model) food. Following DICOM to nifiti-1 conversion using dicom2nii (`http://www.mccauslandcenter.sc.edu/mricro/mricron/dcm2nii.html`), each dataset was subjected to the following preprocessing pipeline carried out in SPM8's batch mode (`http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`). For complete code see, `https://github.com/andsoandso/fmri/tree/master/catreward/spm_m`. Anatomical data was first segmented into white and grey matter regions (Collignon et al., 1995). Based on these segments the parameters necessary for normalization into a standard reference space (T1 MNI-352, at 1 mm, MNI space or short) were

calculated. Normalization had two steps. The first was a Bayesian 12-parameter affine transformation (Ashburner, Neelin, Collins, Evans, & Friston, 1997). The second was a set of nonlinear deformations, using a 1127 parameter discrete cosine transform (Ashburner & Friston, 1999). Anatomical data was then resampled from 1.27 to 1.00 mm^3 using fourth degree β-splines, and finally, using the parameters above, normalized into MNI space.

To correct for the slight head movements that often occurring during scanning, movement regressors for all volumes of the functional data were first calculated (Ashburner & Friston, 1999). No participant moved more than 1.5 mm, so all data was retained. Functional data was then slice-time corrected, using slice 13 (the middle slice from the descending acquisition) as the reference, followed by coregisteration with the pre-processed (native-space) anatomical data, and resampling into 3 mm^3 voxels, again using fourth degree β-splines (Collignon et al., 1995). Functional data was then normalized into MNI space using the anatomically-derived parameters above. Finally, the functional data was spatially smoothed using a 6 mm FWHM Gaussian, though a copy of the unsmoothed data was retained for the ROI analyses (described on p55). Just prior to regression analysis, each voxel's time course was also low-pass filtered using finite impulse response model, with a cutoff at 0.008 Hz (Kruggel, Cramon, & Descombes, 1999). For all whole-brain analyses, the movement regressors were entered into the regression models as covariates, accounting for any head movement. Given the large spatial averages employed in the ROI analyses these weren't motion corrected (Poldrack, 2007).

The best of all possible signals. In fMRI, and in general time-series analysis, there is an intrinsic trade-off between detecting a signal in the presence of noise and estimating the shape of that signal (Dale, 1999; Birn, Cox, & Bandettini, 2002; Liu, 2004). One way to optimize over both these conflicting objectives is to manipulate the trial order in a rapid event-related design (Miezin, Maccotta, Ollinger, Petersen, & Buckner, 2000). One state-of-the-art method for optimizing the trial ordering process is a genetic algorithm which uses two (weighted) loss functions, one for signal detection and one for time-course estimation (Wager & Nichols, 2003). Kao, Mandal, Lazar, and Stufken (2009), improved on Wager's (2003) initial design by adding in a loss function for psychological considerations, greatly improving execution speed and documentation. As a result, Kao et al's (2009) method/code was used to optimize trial orders for part 1 and 2 of the behavioral task (p29), along with the reward category localizer scan (p51).

Mobs of Blobs

All statistical parametric maps (below) were derived from a Random Effects analysis (RFX, or "second-level" in SPM8 jargon), multiple comparison corrected assuming Gaussian Random Fields using the Family Wise Error Rate (FWE) at the $p < 0.05$ level, with a minimum cluster size of 4 voxels (Worsley et al., 1996).

Whole brain activity for the stimulus-response learning portion of the behavioral experiment (i.e., part 2, p29) was examined first by comparing all trials to the baseline (rest) condition. This data is presented in two ways. First is the statistical thresholded image. This contrast map showed significant bilateral activity in the

cerebellum, insula and anterior cingulate ($t(15) = 6.59$, $p < 0.05$; Figure 16). Second is an overlay of the raw t-values, which allows for visual confirmation the observed significant effects were robust and widespread in their respective regions, but also allowed for the analysis of overall and sub-threshold patterns of activity. These raw data suggested near threshold levels of activity in the head of the caudate, ventromedial, dorsolateral frontal cortices as well as (weaker) activity in the occipital lobe (Figure 17). And indeed in a two-way ANOVA looking at that interaction between gains and losses, significant clusters were observed in head and body of caudate, insula, posterior and anterior cingulate with the posterior activation extending into the precuneus, as well as in dorsolateral (i.e middle frontal) PFC, and in ventrolateral PFC (Figure 18; $F(1, 270) = 30.76$, $p < 0.05$). When gains and losses were examined separately, but again compared to rest, both had activity in the same areas as in the combined condition (not shown).

Regions and Models

The right chunks. Following whole-brain analysis, regions of interest were selected using two separate yet related methods. The first employed only regions from the Harvard-Oxford probabilistic anatomical atlas, using the 50% cutoff (Desikan et al., 2006). The second combined anatomical regions with functional clusters isolated using both the data collected during the second half of part 1 (i.e., the "refresher") and from the reward category localizer (p51). Comparisons between them showed the anatomically-limited functional clusters and the entire anatomical regions displayed very similar results. So to limit the complexity of later analyses, and to increase

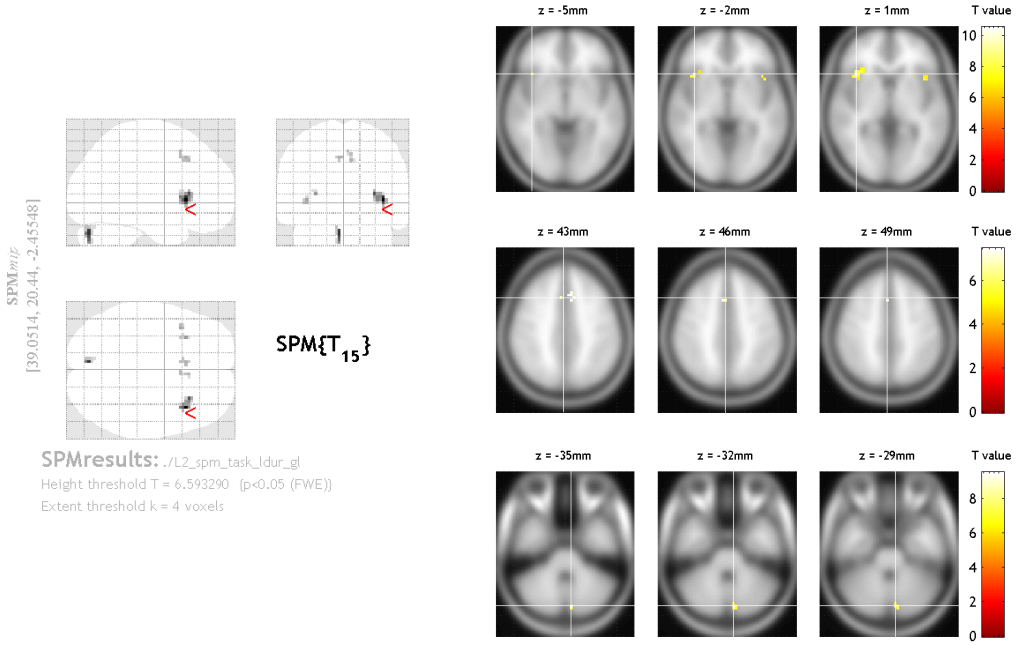

Figure 16. Statistical parametric map for all trials in the stimulus-response learning task (i.e., part 2, p29), compared to the rest period. *Left* is a glass brain, showing all significant clusters. *Right* is a set of axial slices highlighting strong areas of activity overlaid onto the T1 MNI-352 template. Z is the height of the axial slice in MNI space.

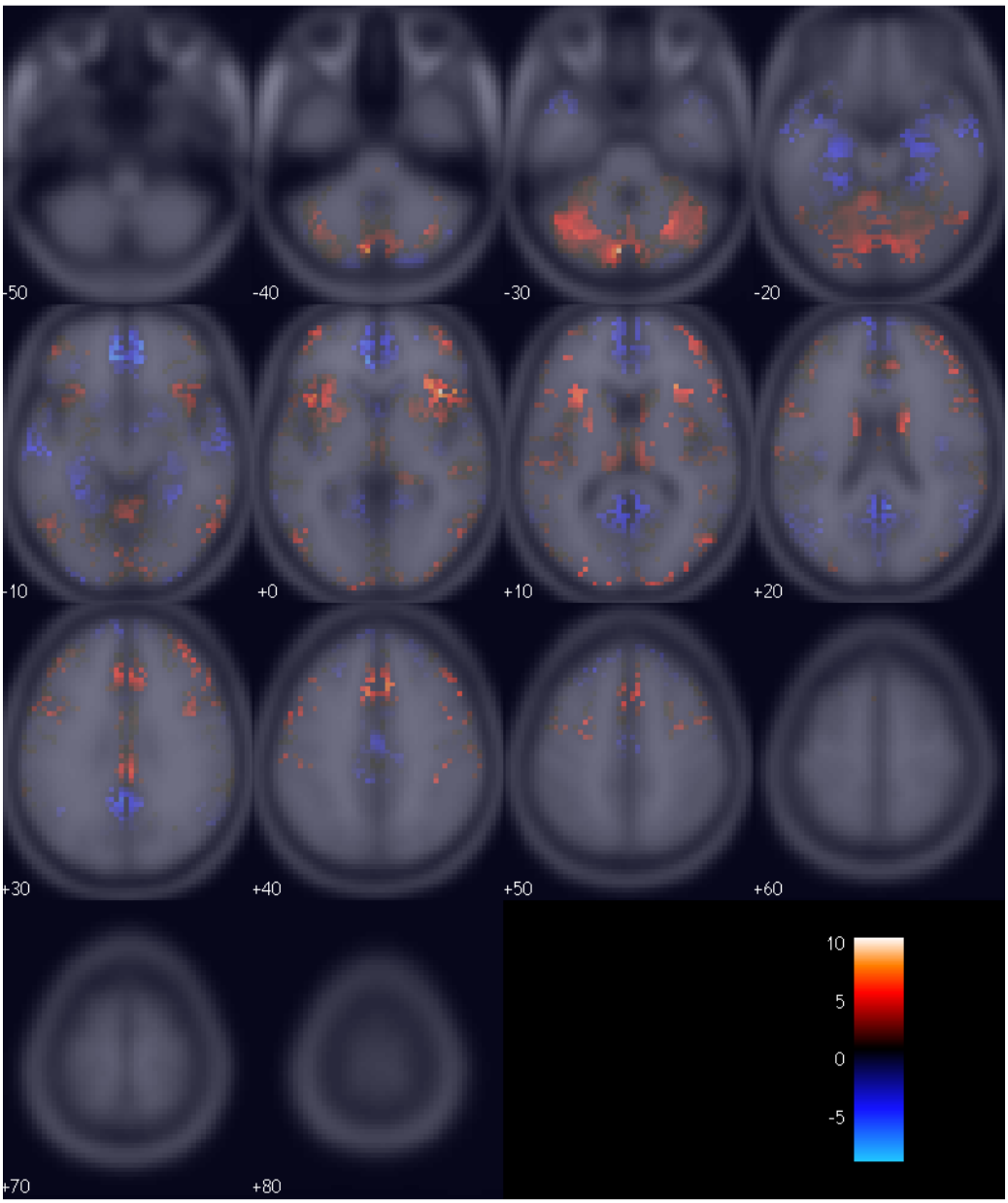

Figure 17. (The t-values for all trials in the stimulus-response learning task (i.e., part 2), compared to the rest period, overlaid onto the T1 MNI-352 template. Each number is the height of the axial slice in MNI space.

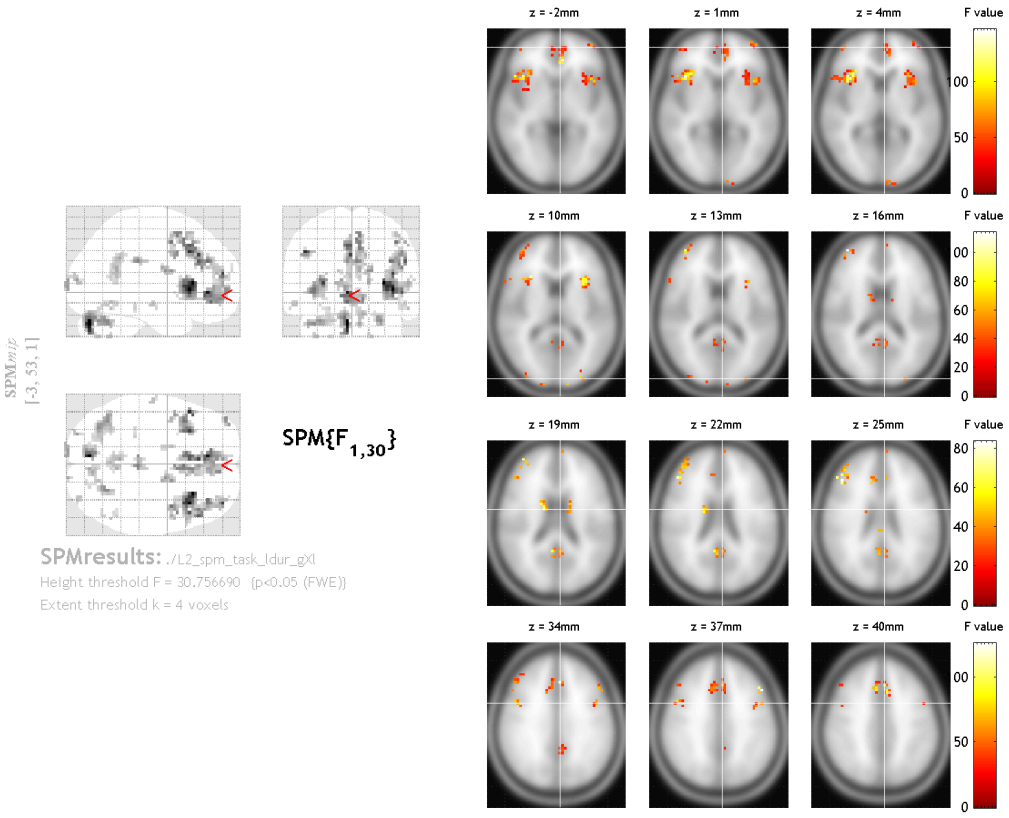

Figure 18. Statistical parametric map for all trials in the stimulus-response learning task (i.e., part 2) examining the interaction between gains and losses. *Left* is a glass brain, showing all significant clusters. *Right* is a set of axial slices highlighting strong areas of activity overlaid onto the T1 MNI-352 template. *Z* is the height of the axial slice in MNI space.

power, functional clusters were discarded in favor of the larger anatomical regions. Most anatomical regions of interest were selected *a priori* based on previous studies of reinforcement and category learning (see the *Introduction* for a review). Left and right subcortical regions of interest were the dorsal caudate, ventral striatum/nucleus accumbens, and putamen. Bilateral cortical areas were the middle frontal cortex (i.e., dorsolateral PFC), frontal medial cortex (which contains ventrolateral PFC), and orbital frontal cortex. Based on the whole-brain maps (p54), regions for the insula, anterior and posterior cingulate (ACC and PCC for short) were included as well. While these is no strong *a priori* hypothesis for the role of these regions may play or may not play, activity in each of these *post hoc* regions is common to human category learning experiments (Lopez-Paniagua & Seger, 2011; Seger et al., 2010; Cincotta & Seger, 2007; Seger & Cincotta, 2006, 2005).

A way to(o) many. There are 6 models under evaluation: the three kinds of similarity adjustment ("none", "exp", and "gauss") multiplied by the two possible reward codes ("acc" and "gl"), with the two terms of interest (i.e., value and the reward prediction error), that is 12 comparisons. The were also a number of *a priori* confounds to the signals of interest including the similarity metrics, the reward codes, and the grating parameters, bringing the total to 23. As the models are not nested[12] and therefore not amenable to F-tests – the common statistical way to compare model fits – an alternative approach was called for. Further complicating the issue was the fact that each of the models is covariate, if not collinear, with

[12]Often defined by whether or not two models can be made identical by adding or subtracting parameters (Forster, 2000)

the others. To top it off, none of the three similarity-adjustments are statistically independent; reinforcement learning can be viewed as a regression of the reward code onto behavioral choices. All these factors combined would make statistical testing difficult, to say the least. But fortunately finding *the* best model is not the goal.

The latest recordings of phasic (i.e., reward prediction) activity in the VTA/SNc suggests a complicated reward and prediction error coding scheme (see p7), wherein several separate sets of calculations may be carried out independently (H. Kim et al., 2006; Matsumoto & Hikosaka, 2009; K. S. Smith et al., 2011). The observed BOLD signal is then an aggregate of these many activities. It is possible, even likely, then that more than one of the models is correct making null hypothesis tests an incorrect choice. Model selection is the right choice.

Model selection is the process of finding a *family* of models that best predict a given dataset (Rao, Wu, Konishi, & Mukerjee, 2001). Most techniques try to wisely balance parsimony with increasing fit (i.e., solving the bias versus variance dilemma (Geman, Bienenstock, & Doursat, 1)). Unfortunately most model selection techniques require assumptions the models cannot meet (e.g., statistical independence). The few that can tend to be complex recent statistical inventions. Rather than navigate those troubled and unproven waters, a simpler approach was taken. Each model was independently examined and ranked, in an approach loosely similar to model averaging (Forster, 2000).

An AIC score (Akaike Information Criterion (Akaike, 1974)) was assigned to each of the models/codes for every participant and region of interest. The absolute

AIC score across participants is not however meaningful. Only the relative values are of interest (Wagenmakers & Farrell, 2004). As a result, individual's scores were normalized and ranked by subtracting the best (lowest) score from each (Anderson, Burnham, & Thompson, 2000). The normalized set was then transformed to Akaike Weights, a way to easily compare the conditional probabilities of each model being true (Wagenmakers & Farrell, 2004). The Akaike Weights were then averaged across participants for each model and region of interest.

Information on information. AIC is a measure of loss; how much information is lost by substituting the model for the true distribution, i.e., the data. The lower the AIC score, the better the model. Unlike null hypothesis tests and Bayesian measures, AIC-based methods do not seek to find *a* truth, but instead serve to rank models. AIC offers then only relative insight, and is unable to make any claims about absolute significance. Significance is a separate question, one to be returned to later. Besides this limitation, AIC has some substantial advantages. Five are reviewed below.

One, unlike maximum-likelihood, AIC is designed to be a parsimonious score. It penalizes for additional parameters. It may therefore choose a worse model (as measured by likelihood or mean squared error) over a better but more complex one. This is the essence of Occam's razor[13].

Two, it fits with the process of science. When designing an experiment it is rare that there are only two possible outcomes, instead typically there are several competing hypothesis, some of which may not be mutually exclusive. AIC's focus

[13]Famously and pithily expressed as, "Entities are not to be multiplied beyond necessity".

on relative differences and evidential weights meshes perfectly with the reality of multiple working hypotheses (Burnham, 2004).

Three, truth can remain elusive. A common alternative to AIC is BIC, the Bayesian Information Criterion. Like AIC, BIC is derived from the log-likelihood of a model, however its derivation requires a rather strict (and often unrealistic) assumption – that the true model is among the candidates (Forster, 2000). And while it may be philosophically debatable whether any mathematical model can *completely* describe reality, in this study the models are incomplete. As, one, the human reinforcement learning literature contains several recent theoretically unaccounted for findings and, two, there are theoretical developments not include here to keep the models tractable (see the *Introduction* for a review).

Four, AIC values are easily interpretable once they're transformed to Akaike Likelihoods or Weights[14]. The likelihood is, as you would expect, simply the likelihood the model is correct (based on the information loss associated with it), while the Akaike Weights are normalized likelihoods. As the Weights sum to one, the conditional likelihood of one model compared to another is just the ratio of their weights (Burnham, 2004). For example, the conditional likelihood of model A over model B is just w_A/w_B. That is, the likelihoods and Akaike Weights are intrinsically measures of effect size (Anderson et al., 2000; Forster, 2000). Despite the fact that it is often used to express the likelihood of correctly rejecting the null hypothesis,

[14]Likelihood for model k among K working hypotheses/models is given by $L_k = e^{-0.5(AIC_k - min_K(AIC))}$, which is then normalized, becoming an Akaike Weight by $w_k = L_k / \sum_{k=1}^{K} L_k$ (Burnham, 2004).

the p value is not a measure of effect, as p is contingent not just on effect size but on sample number.

Five, AIC has a history with models of categorization. McKinley and Nosofsky (1996); Maddox and Bohil (2001), among several others, used AIC to compare behavioral results to several alternative models of categorization.

F-them. AIC ranks offer no information about significance, in the familiar null hypothesis sense, or about the absolute fit of the model. Both of these were addressed in a series of F-tests run prior to AIC analysis. These (fixed-effect, across participant) omnibus tests asked whether the total set of regression parameters for each linear model (described below) could explain the BOLD time series better than chance (i.e could the null hypothesis (of 0) be rejected). Keeping with recommendations of Burnham (2004); Forster (2000), who argue that as AIC and significance tests are so dissimilar that direct comparison/interaction between them will be at best misleading, the models are not discarded based on significance. All models are retained, and later AIC ranked. The F-tests are a separate measure whose results are integrated during interpretation, not during model selection.

As is discussed in the results (for example see p69), many of the models all are significant by these omnibus F-tests, which might at first be rather shocking. Many of the models and other regressors, including the simplistic "boxcar", make very different predictions about the BOLD response. One would, therefore, expect only a few to be significant. However this intuitive prediction is false, as each predictor is convolved with a haemodynamic response function which spans more than 20

seconds (p64). As a result, each single time point prediction is "smeared" across that temporal distance. Unpublished Monte-Carlo simulations on the effect of this "smearing", confirm that null hypothesis tests, such as the F-tests here, are a poor metric of model quality and specificity, i.e., the likelihood the true model will be picked over some (related) alternatives.

Code, BOLD, and models.. A total of 23 models were compared for each of the 12 regions of interest for each of the 16 subjects, 4416 comparisons in total. Each of the models is described below (Table 1). In general, a time-series (e.g the reward prediction error for each trial or the similarity for that trial's outcome) was convolved with a "canonical" haemodynamic response function, a mixture of gamma functions that serves as a parsimonious estimate of the (instantaneous) BOLD response (Friston et al., 1998). The convolved series was then low-pass filtered, matching the treatment of the BOLD data (p52). Each convolved and filtered model was then regressed onto the BOLD response for each participant's region of interest, retaining all parameters and fit measures inside subject-level HDF5 files. The HDF5 format offers high performance read/write operations, and widespread support across several scientific programming languages (`http://www.hdfgroup.org/HDF5/`).

No available fMRI analysis package returns AIC scores (or measures that could be converted to such) and none allow for the efficient (i.e programmatic) analysis of many competing computational models. So a region of interest focused fMRI analysis tool was created in Python (v2.7.1) to meet those two needs. This module, simply named "roi", has since been release under the BSD license and

is available for download at `https://github.com/andsoandso/roi`. It relies on the nibabel library to read the nifiti-1 files (v1.2.0; `http://nipy.org/nibabel`), nitime for time-series analysis, (v0.4; `http://nipy.sourceforge.net/nitime/`) Numpy for generic numerical work (v1.6.1; `http://numpy.scipy.org/`), with the GLS function from the scikits.statsmodels module handling the regressions (v0.40; `http://statsmodels.sourceforge.net/`). Model-to-BOLD fit parameters, as well as other useful metadata, was then extracted and stored in text files suitable for importing into R (v2.15.1; `http://www.r-project.org/`). All plotting and model ranking (as well as the F-tests) were carried out in R. For complete BSD licensed code see, `https://github.com/andsoandso/fmri/tree/master/catreward/roi/results`.

Our kinds of models. To ease visualization and analysis each of the models was classified into one of 5 families. Family one, denoted "boxcar", was identical to that first used in the whole-brain analysis (p54) – all trials versus the rest condition. This is a univariate time-series that predicts no trial-specific effects; No matter the task the brain, thus the BOLD response, just flicks on then off. It serves as a useful standard against which to compare the model-based regressors. The next two families were controls (i.e., *a priori* covariates). The reward codes, both raw and similarity adjusted, were in one family ("control_reward") and in the other were the similarity metrics and grating parameters ("control_similarity"). The fourth family contained all the reward prediction errors ("rpe"). The fifth contained all value estimates ("value").

Table 1:: All models, their designations (Codes), families, and descriptions.

Number	Code	Family	Description
1	0_1	boxcar	The simplest model, a univariate analysis of all conditions.
2	acc	control_reward	Behavioral accuracy.
3	acc_exp	control_reward	Behavioral accuracy, diminished by (exponential) similarity.
4	acc_gauss	control_reward	Behavioral accuracy, diminished by (Gaussian) similarity.
5	gl	control_reward	Gains and losses.
6	gl_exp	control_reward	Gains and losses, diminished by (exponential) similarity.
7	gl_gauss	control_reward	Gains and losses, diminished by (Gaussian) similarity.
8	rpe_acc	rpe	Reward prediction error - derived from accuracy.
9	rpe_acc_exp	rpe	Reward prediction error - derived from accuracy diminished by (exponential) similarity.

10	rpe_acc_gauss	rpe	Reward prediction error - derived from accuracy diminished by (Gaussian) similarity.
11	value_acc	value	Value - derived from accuracy.
12	value_acc_exp	value	Value - derived from accuracy diminished by (exponential) similarity.
13	value_acc_gauss	value	Value - derived from accuracy diminished by (Gaussian) similarity.
14	rpe_gl	rpe	Reward prediction error - derived from gains and loses.
15	rpe_gl_exp	rpe	Reward prediction error - derived from gains and losses diminished by (exponential) similarity.
16	rpe_gl_gauss	rpe	Reward prediction error - derived from gains and losses diminished by (Gaussian) similarity.
17	value_gl	value	Value - derived from gains and losses.

18	value_gl_exp	value	Value - derived from gains and losses diminished by (exponential) similarity.
19	value_gl_gauss	value	Value - derived from gains and losses diminished by (Gaussian) similarity.
20	exp	control_similarity	Outcome similarity (exponential).
21	gauss	control_similarity	Outcome similarity (Gaussian).
22	angle	control_similarity	Grating angle parameter.
23	width	control_similarity	Grating width parameter.

Model Results

The many results are dicussed, first by subcortical areas then moving on to the cortical. The general analysis strategy was to first find the top family, indicated by the largest family-average Akaike Weight. Then the next highest scoring to family was examined to see if it was close to the top (i.e., $\leq$1.5 times as likely). If it was both, families were included. The next step examined the relative likelihood of each model in the top family/families. Within-family models that were about $\geq$1.5 times more likely then their neighbor were dubbed "substantively more informative".

Like the significance thresholded in null hypothesis tests this $\geq$1.5 is an arbitrary threshold. However in order discuss and interpret these results a line must be drawn between meaningful and not, and $\geq$1.5 is a good minimum cutoff (Anderson et al., 2000; Forster, 2000). As was stated at the outset, more than one model may be right. Thus the threshold was treated as a loose cutoff. To get a sense of overall model quality, I the likelihood of the best model over the boxcar (i.e., the non-parametric standard) was calculated. Finally all models, not just the top family, were assesed for any outliers that may have scored well despite their families overall poor performance.

As this was the first attempt to AIC-rank models of fMRI data, and while much thought and research was put into the above scheme, it may be flawed. It is also arbitrary (beyond the $\geq$1.5 cutoff); Why not discuss the top 3, or 4 families, or even just include them all? To attempt then to minimize the effect of these arbitrary, but necessary, decisions the complete set of models (and F-tests) are included for every region of interest.

From up high. For eight of the twelve regions of interest the "rpe" family scored highest. Of these eight, five were best described by "rpe_acc_gauss". The next best family was "control_similarity" with 3 regions, followed by "boxcar" with 1. Notably, "value" was not the most informative model family for any region of interest, and indeed the one region (ACC) for which it was second, "rpe" was 1.8 times more likely.

Under cortical. In the dorsal caudate (Figure 19), only the "rpe" family offered a more informative fit that the "boxcar", being 2.61 times more likely in the left and 2.85 in the right (abbreviated as left/right: 2.61/2.85 from here on). Bilaterally, and using the "acc" coding scheme, the Gaussian similarity-adjusted model (i.e., "rpe_acc_gauss") was substantively more informative than either unadjusted model ("rpe_acc" – 1.45/1.54 or "rpe_gl" 1.82/1.70). Surprisingly, given its similarity to the Gaussian adjustment, "rpe_acc_exp" scored no better than the unadjusted models (above). In what will become a reoccurring theme when examining the F-tests, all models were significant bilaterally in the dorsal caudate (Figure 20). And while the F-values themselves to some degree mimic the patterns of the Akaike Weights, it would not be possible to reliably disassociate them given the slight relative differences.

Compared to the "boxcar", the putamen was also best described by the "rpe" family (1.53/2.31). However compared to caudate the putamen displayed markedly different within-family activity (Figure 19 compared to 21). The "rpe_acc" model was more substantively more likely (1.67/1.66) than the next highest ranking similarity model (i.e., "rpe_acc_gauss"). However due to the marginal bilateral significance (Figure 22), this interesting reversal must be viewed cautiously. The bilateral consistency in the Akaike Weights does offer some room for optimism (Figure 21, specifically refering to the consistency and relative strength of "rpe_acc").

The right and left halves of the nucleus accumbens, the ventral portion of the striatum, were quite divergent in their fits (Figure 23). However both ranked the "control_similarity" family as the most informative, however this family was only

about 1.10 times more likely than the next family ("rpe"), which itself was not substantively better then its neighbor, and so on. So while there is strong evidence that the top models, "angle" in left (3.07) and "rpe_gl" (3.06) in the right, are better than "boxcar", the overall bilateral heterogeneity, weak family effects, combined with by-and-large non-significant outcomes on the left half, and weak F-values on the right (Figure 24), suggest this region was not strongly activated by the task. Further analysis is therefore futile.

On that thinking sheet. The insula was the one region, both cortically and subcortically, to be nearly equally well described both by the "acc" and "gl" reward codes (Figure 25). In the top ranking "rpe" family (which was 1.37 times more likely than its neighbor "control_similarity", and 2.31 more likely than the "boxcar") within-family models showed divergent patterns based on the code. The "acc", "rpe_acc" and "rpe_acc_gauss" models dominated "rpe_acc_exp" (around 1.4). The "gl" code had the opposite effect, "rpe_gl_exp" dominated "rpe_gl" and "rpe_gl_gauss" by a somewhat similar amount (1.20). The strong overall significance of all models (Figure 26) suggests these relative rankings may reflect truth; like the dorsal caudate the F-values the have same patterns as Akaike Weights.

Both the ACC and PCC displayed very similar rankings of their Akaike Weights (compare Figures 27 to 29), so they'll be discussed as one. Again the "rpe" family dominated (respectively 2.07 and 1.77 times more likely compared to the nearest neighbor, 2.86 and 2.88 times more likely then the "boxcar"). Unlike the caudate and insula, the 2 similarity-adjustment models ("rpe_acc_gauss" and "rpe_acc_exp")

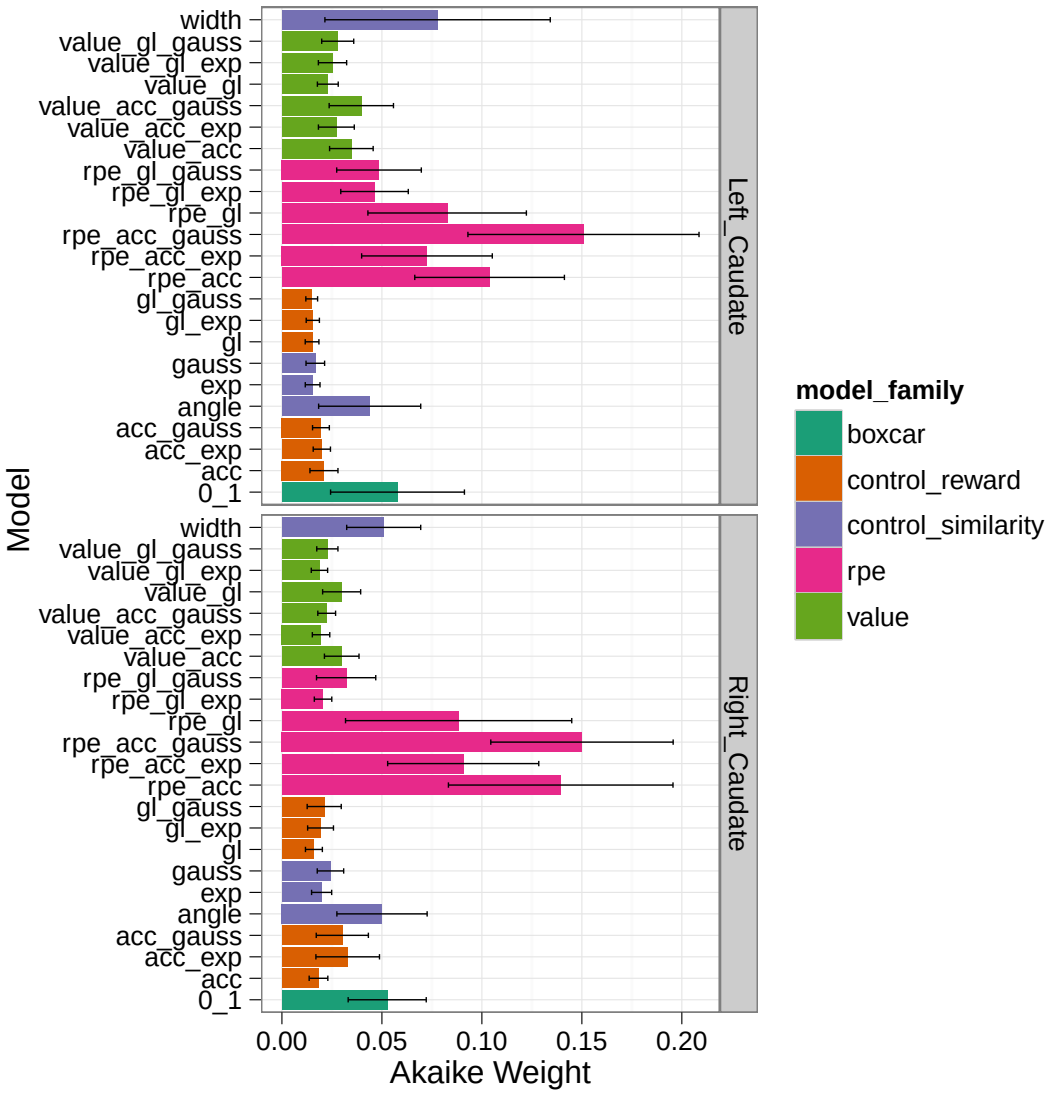

Figure 19. Dorsal caudate (left and right) – Akaike Weights for all models. Colors indicate model family (see p64 for details). Bars represent standard errors.

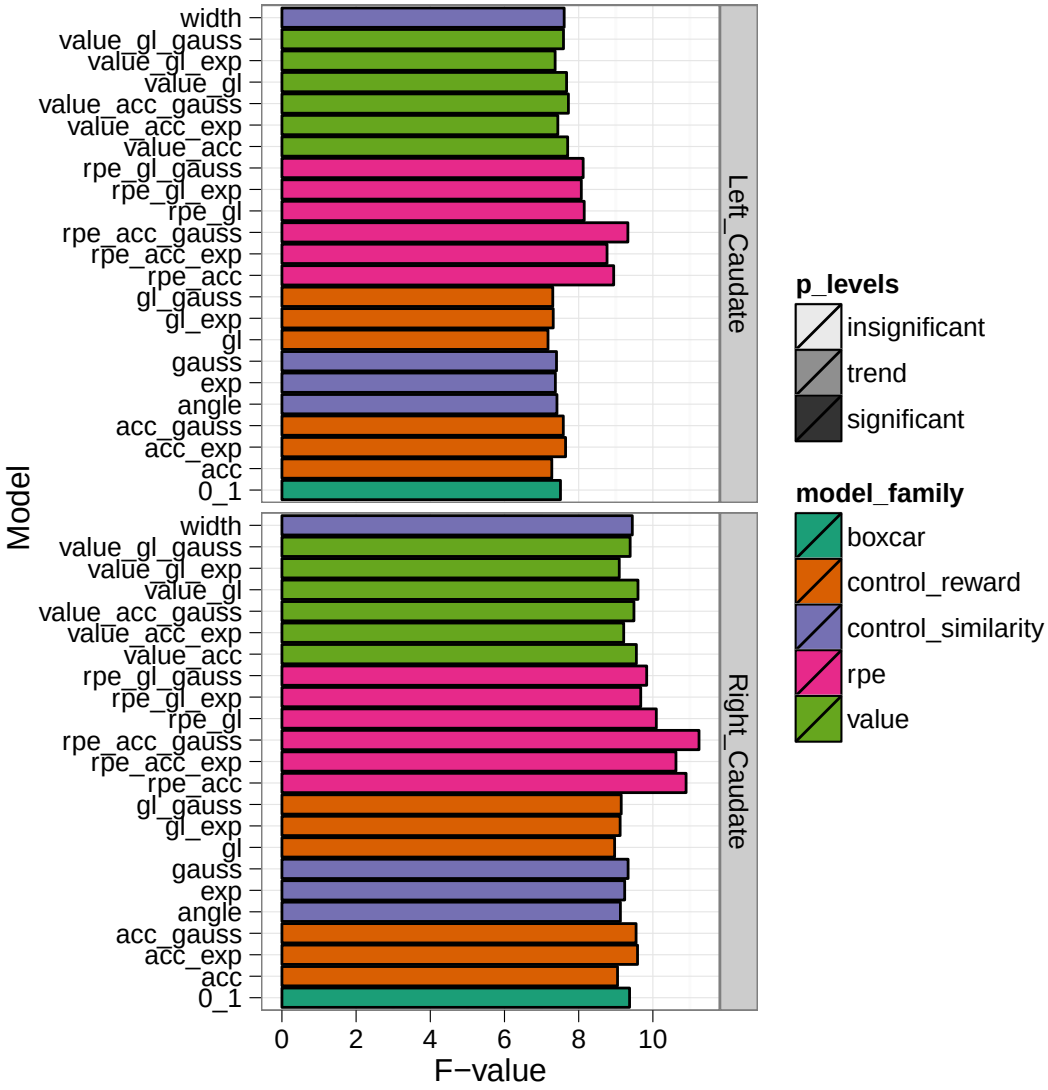

Figure 20. Dorsal caudate (left and right) – F-values for all models. Significance-level is denoted by the saturation, where the $p < 0.05$ level is significant, and trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p64 for details).

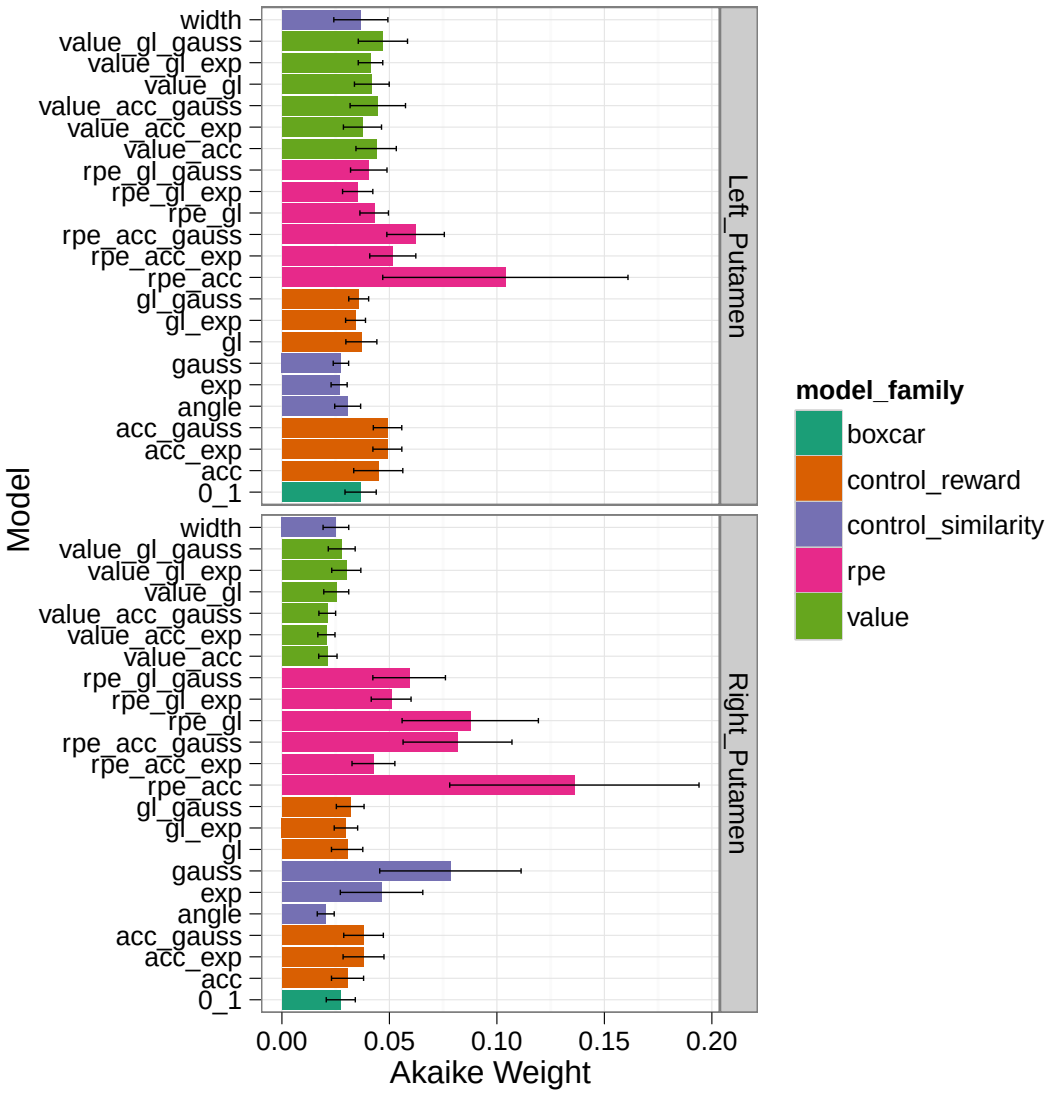

Figure 21. Putamen (left and right) – Akaike Weights for all models. Colors indicate model family (see p64 for details). Bars represent standard errors.

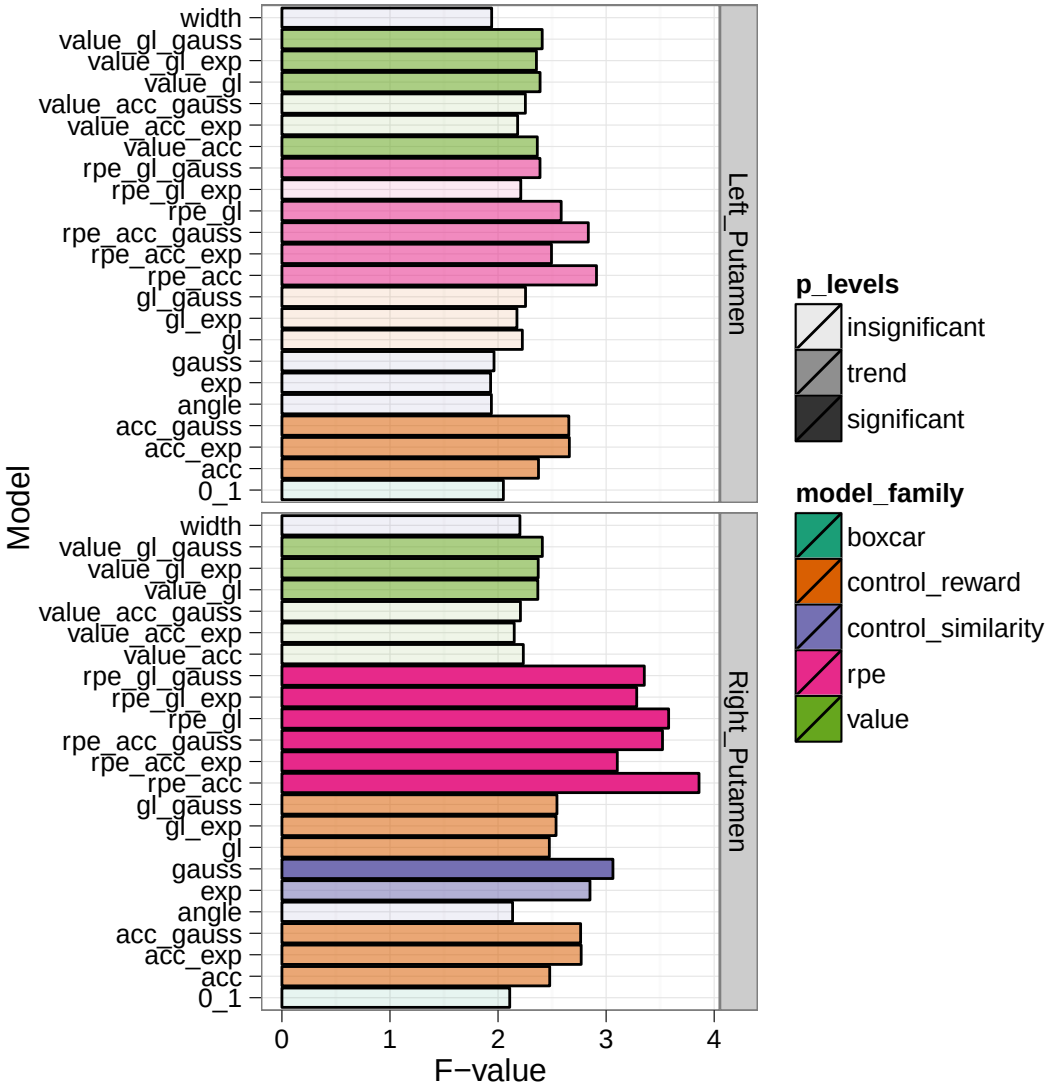

Figure 22. Putamen (left and right) – F-values for all models. Significance-level is denoted by the saturation, where the $p < 0.05$ level is significant, and trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p64 for details).

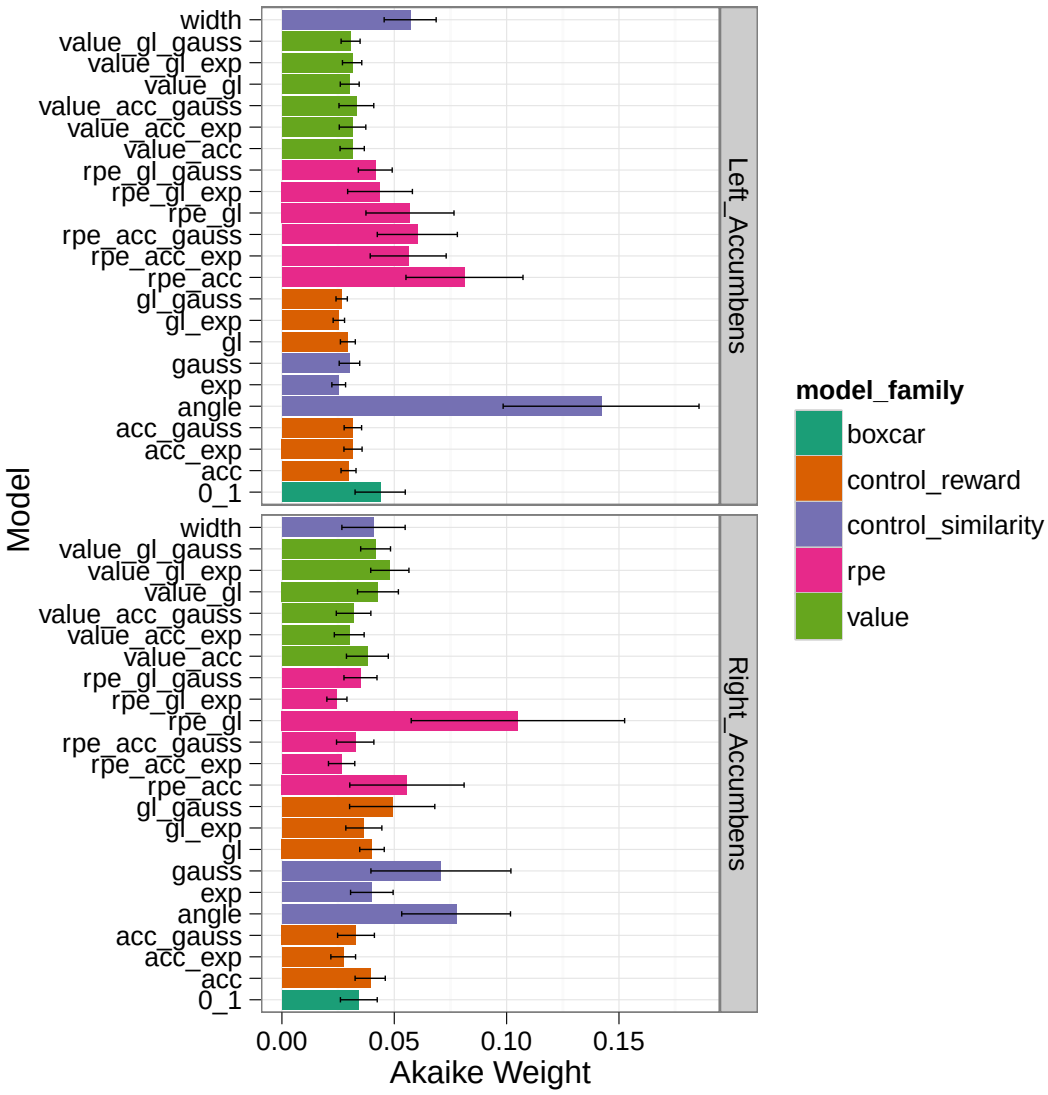

Figure 23. Nucleus Accumbens (left and right) – Akaike Weights for all models. Colors indicate model family (see p64 for details). Bars represent standard errors.

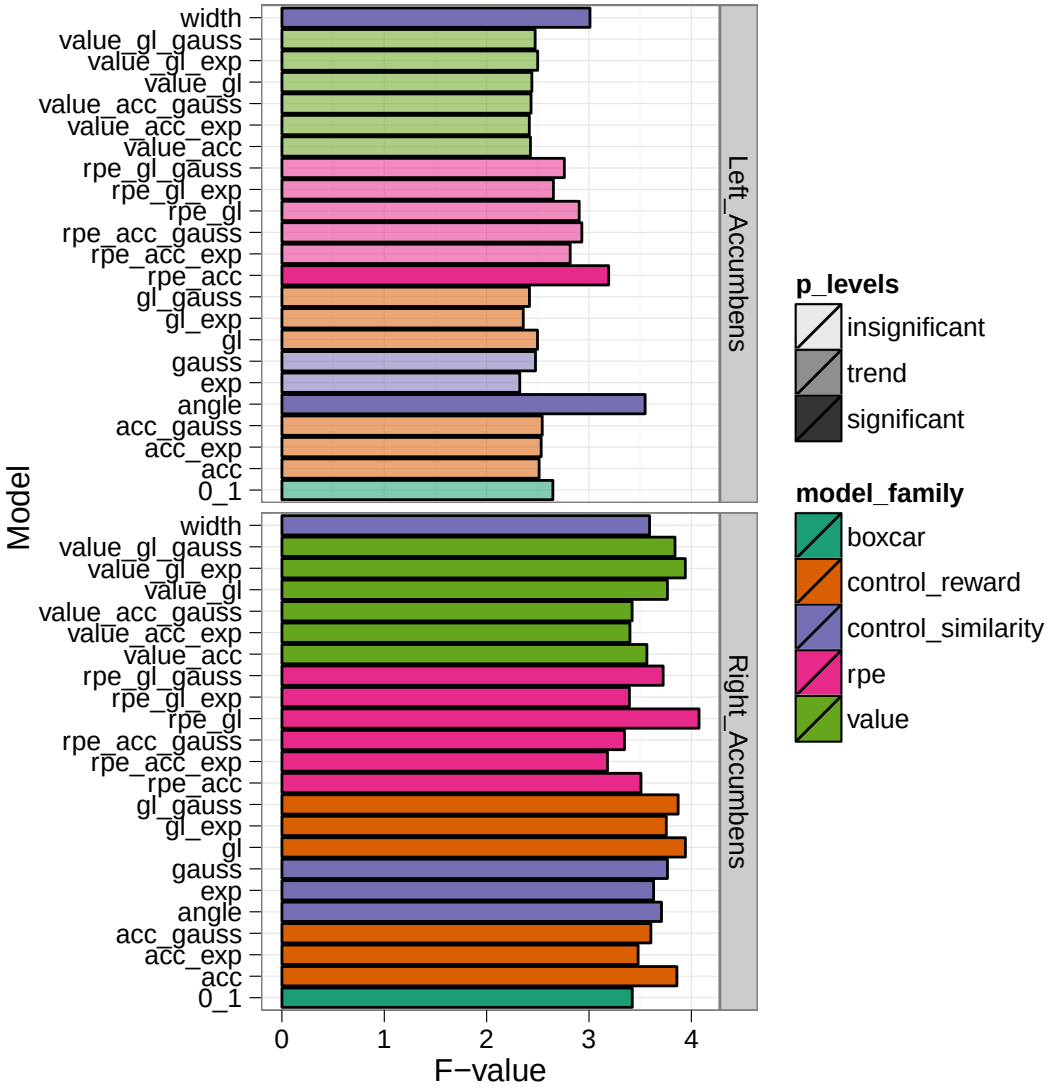

Figure 24. Nucleus accumbens (left and right) – F-values for all models. Significance-level is denoted by the saturation, where the $p < 0.05$ level is significant, and trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p64 for details).

were consistently more informative than the reward unadjusted ("rpe_acc"). Looking at the F-tests, both regions, especially in the "rpe" family were, were reasonably significant (Figure 28 and 30).

Like the orbital frontal (below), "boxcar" was the most informative family for the middle frontal cortex (Figure 35), however model-wise "rpe_acc" was 2.07 times more likely. The strong F-values for all models, and "rpe_acc" especially (the largest observed for all models and regions; Figure 36), lend strong support then for selecting "rpe_acc" as the sole best explanation of this regions activity.

Both the frontal medial (ventrolateral) and orbital frontal results are difficult to interpret, but for different reasons. Nearly all families, and models, in the in frontal medial cortex ranked as substantively more likely that the "boxcar" (ranging, at the family level, from 2.13 for "control_similarity" to 1.75 for "value", with higher scores for individual models). However none of the models were substantively (or even slightly) more likely than any of their neighbors. So while, as measured by the F-tests, there was significant activity in medial frontal (Figure 34) it is not well accounted for by any of the candidate models. Orbital frontal cortex though has the opposite problem. The "boxcar" had the largest Akaike Weight (Figure 31). However at the model-level "rpe_acc_gauss" and "rpe_acc" both score slightly better (1.32 and 1.34, respectively). However these slight increase are weak evidence when the alternative is that nothing has changed from trial-to-trial (i.e., the "boxcar" model). Additionally as ACC and OFC are tightly functionally interconnected (Rudebeck et al., 2008), these weak likelihoods may be carryover from the strong "rpe" signals observed in the ACC (Figure 27).

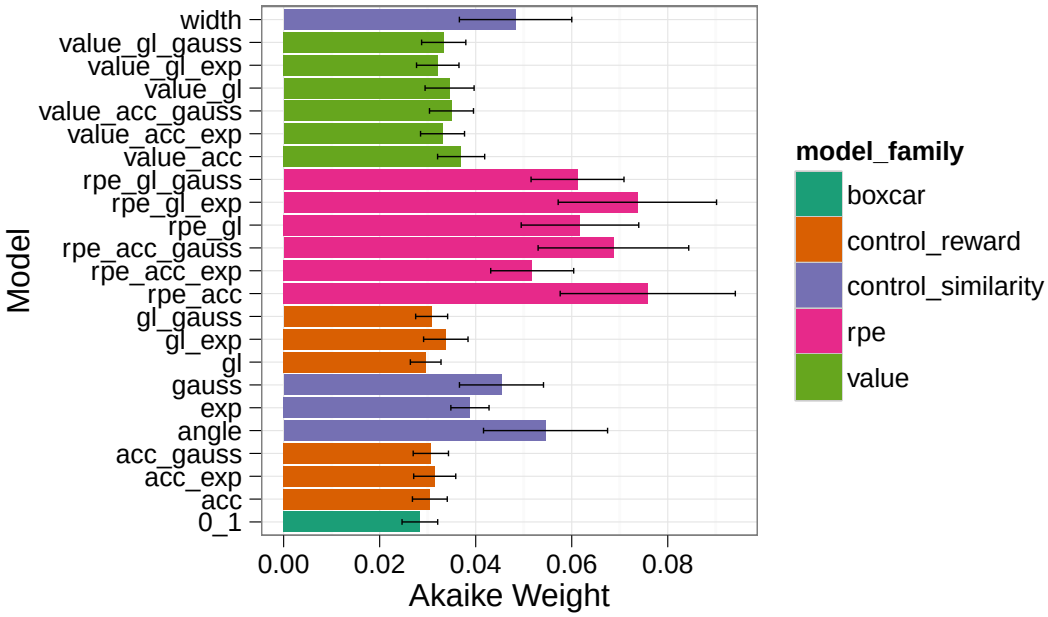

Figure 25. Insula – Akaike Weights for all models. Colors indicate model family (see p64 for details). Bars represent standard errors.

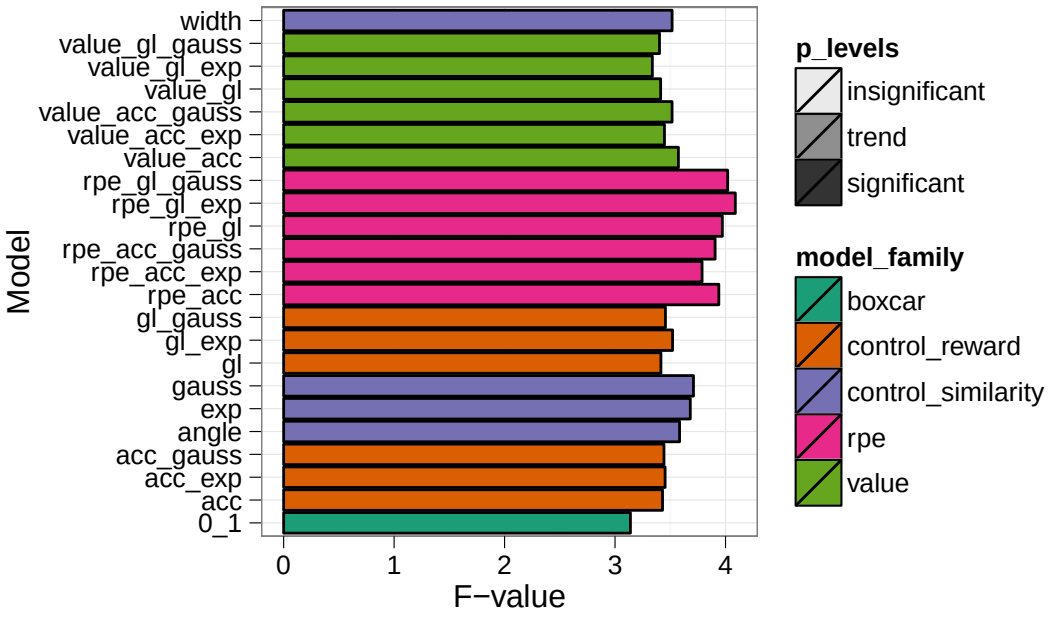

Figure 26. Insula – F-values for all models. Significance-level is denoted by the saturation, where the $p < 0.05$ level is significant, and trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p64 for details).

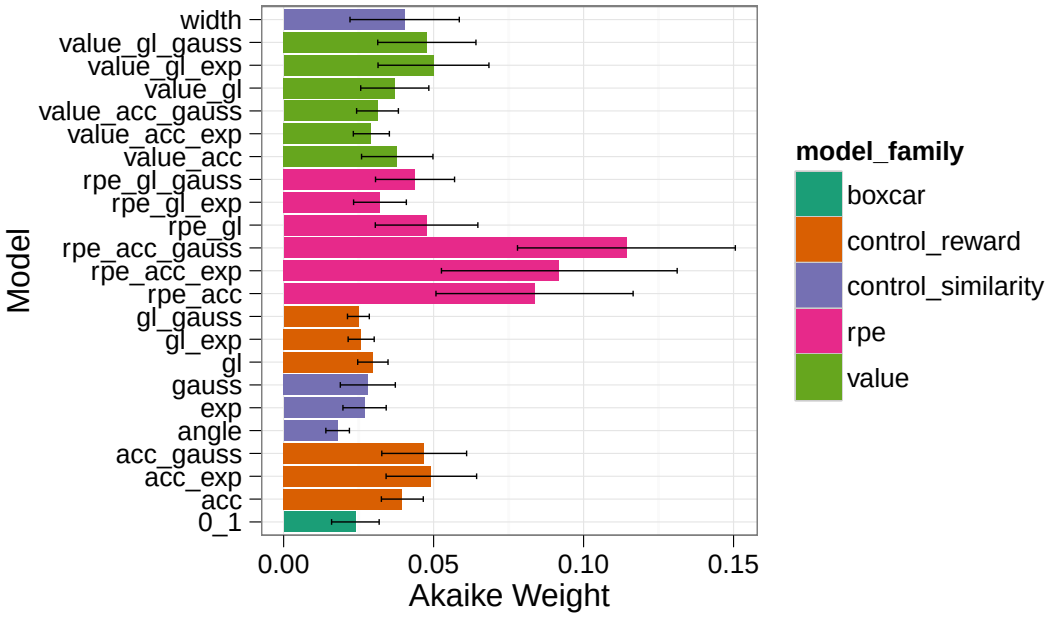

Figure 27. ACC – Akaike Weights for all models. Colors indicate model family (see p64 for details). Bars represent standard errors.

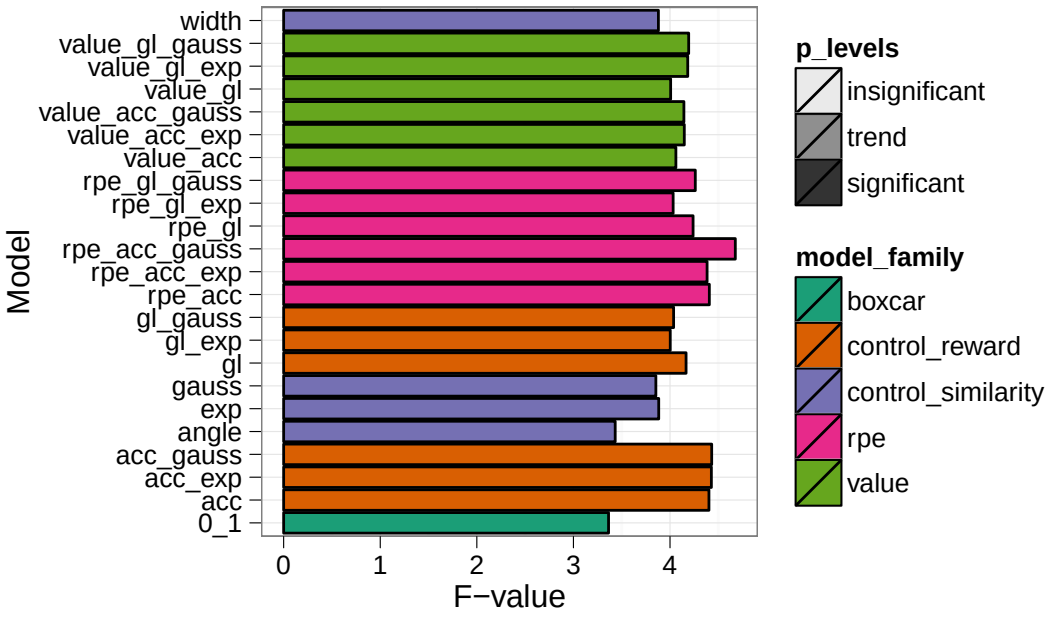

Figure 28. ACC – F-values for all models. Significance-level is denoted by the saturation, where the $p < 0.05$ level is significant, and trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p64 for details).

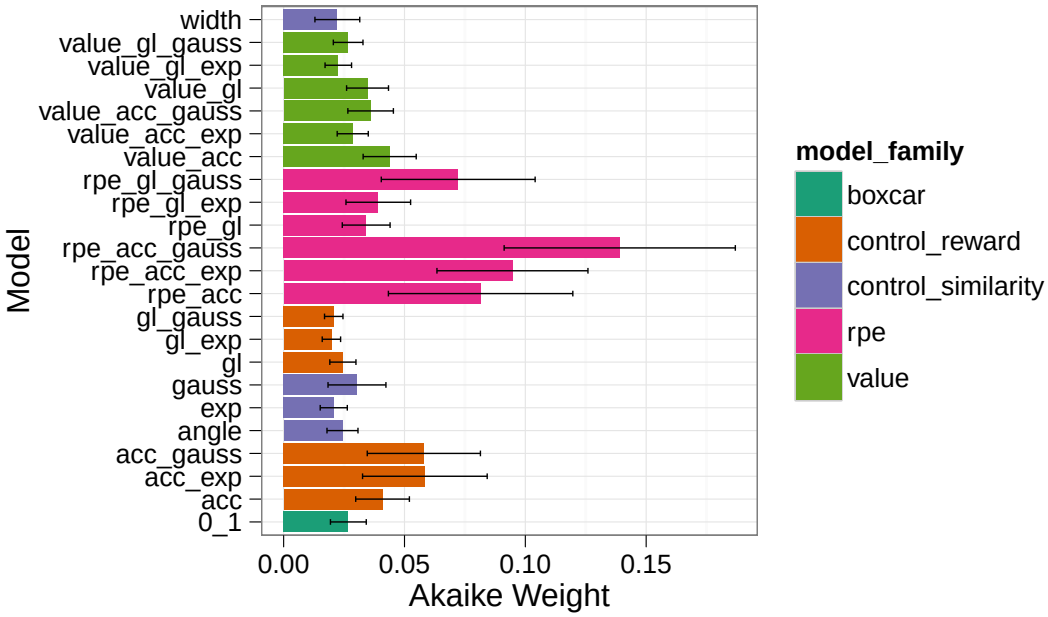

Figure 29. PCC – Akaike Weights for all models. Colors indicate model family (see p64 for details). Bars represent standard errors.

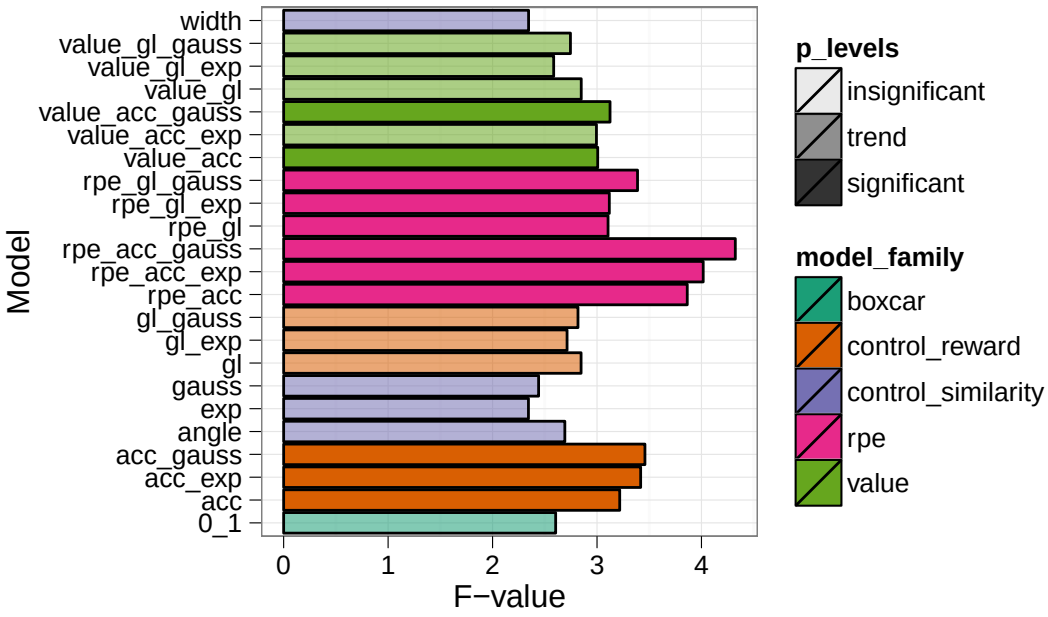

Figure 30. PCC – F-values for all models. Significance-level is denoted by the saturation, where the $p < 0.05$ level is significant, and trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p64 for details).

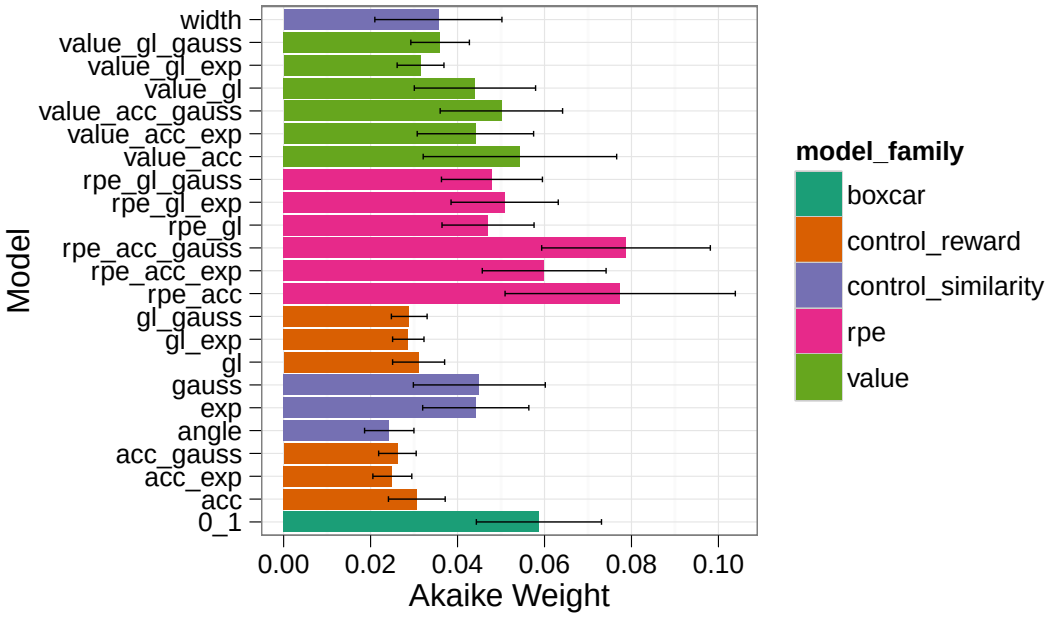

Figure 31. Orbital frontal cortex – Akaike Weights for all models. Colors indicate model family (see p64 for details). Bars represent standard errors.

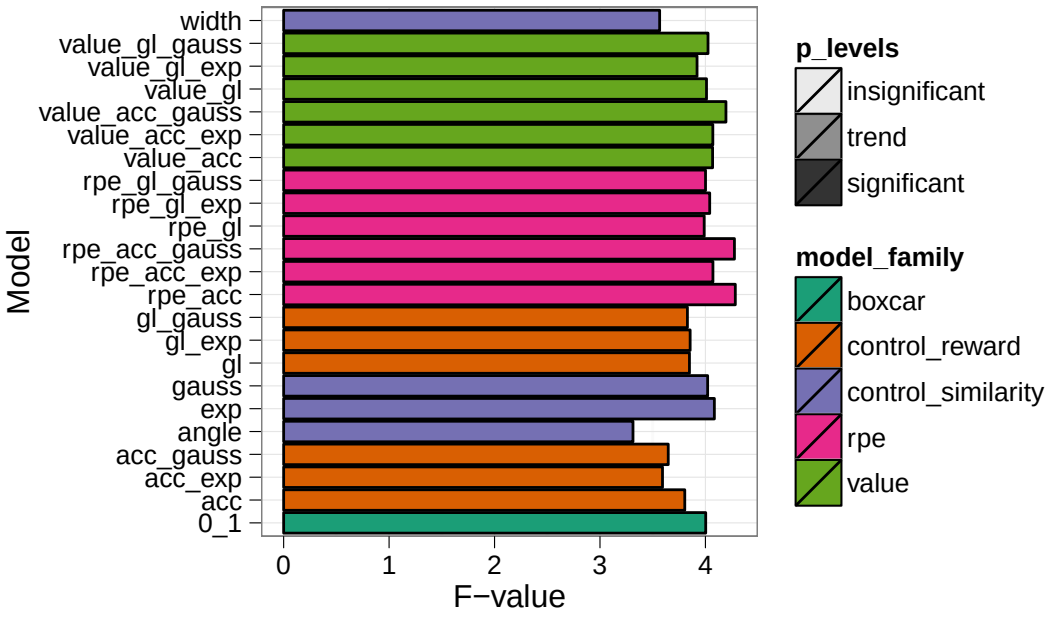

Figure 32. Orbital frontal cortex – F-values for all models. Significance-level is denoted by the saturation, where the $p < 0.05$ level is significant, and trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p64 for details).

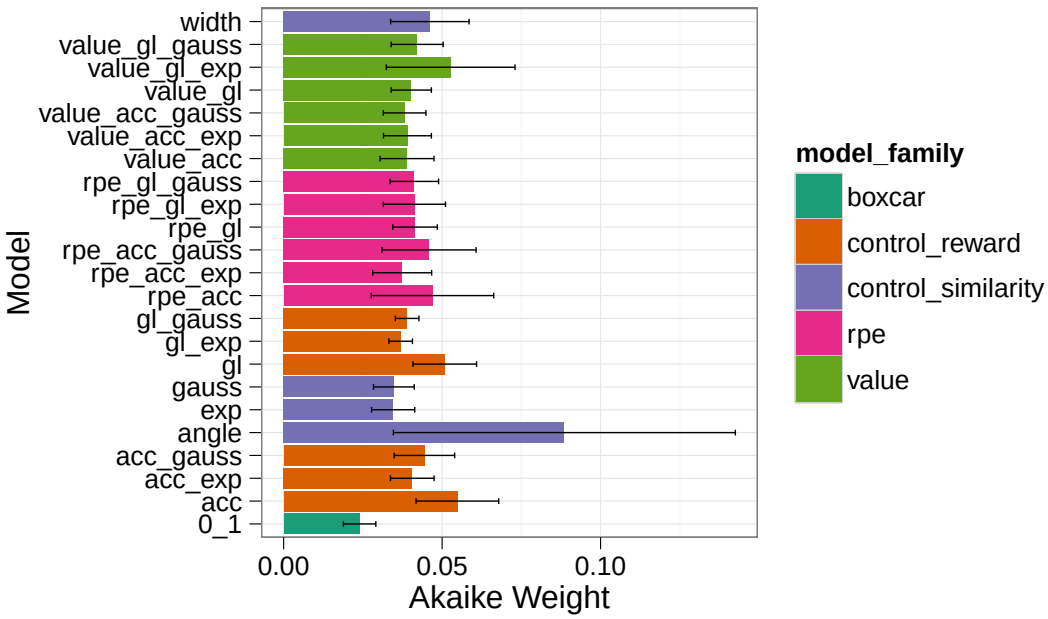

Figure 33. Frontal (ventral)medial PFC – Akaike Weights for all models. Colors indicate model family (see p64 for details). Bars represent standard errors.

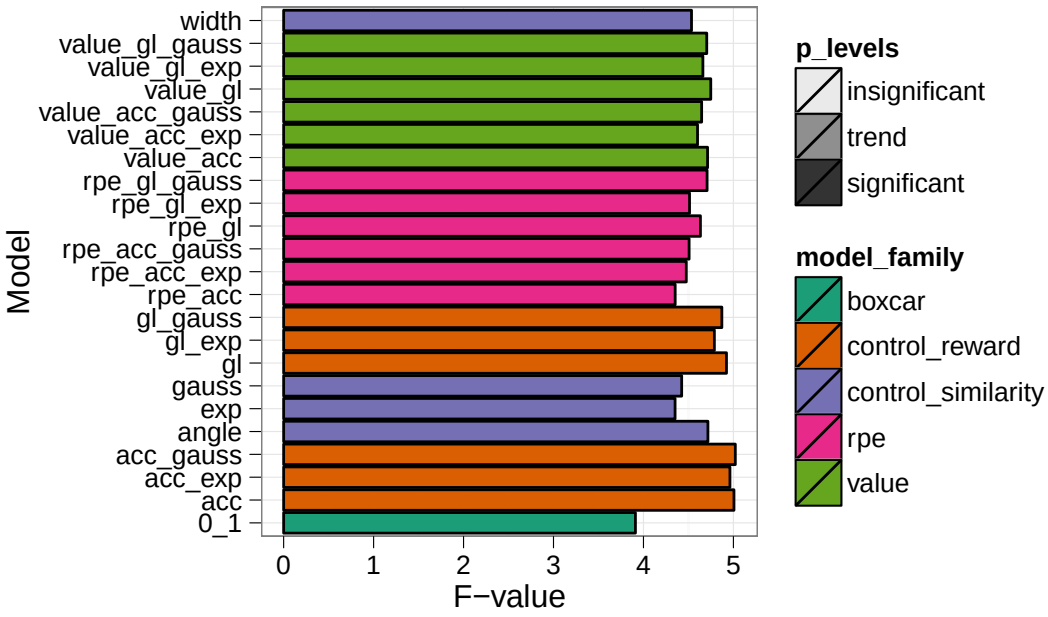

Figure 34. Frontal (ventral)medial PFC – F-values for all models. Significance-level is denoted by the saturation, where the $p < 0.05$ level is significant, and trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p64 for details).

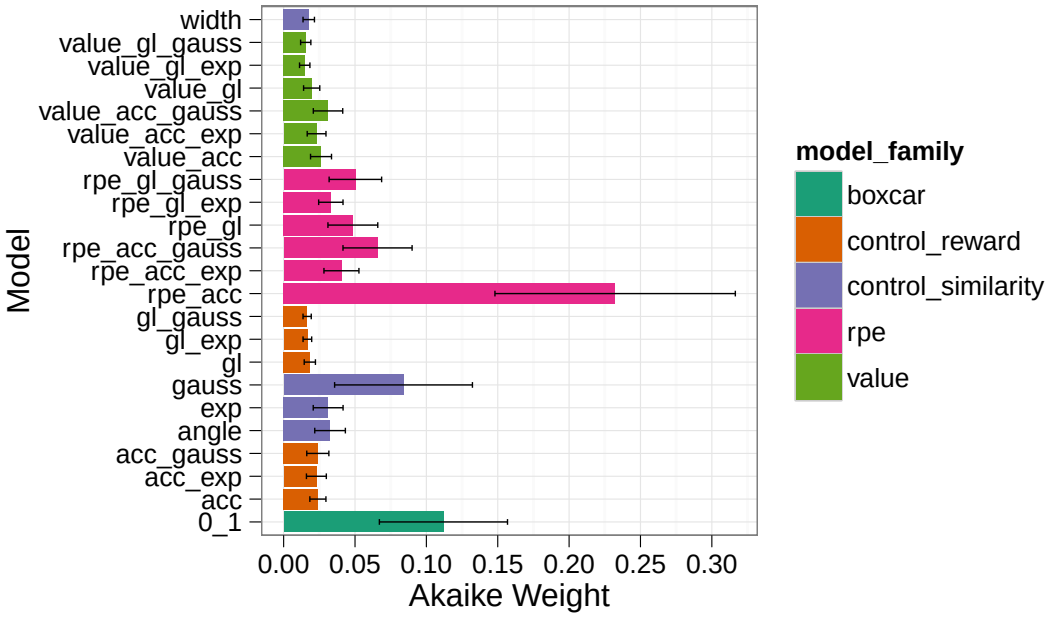

Figure 35. Middle frontal (dorsolateral) PFC – Akaike Weights for all models. Colors indicate model family (see p64 for details). Bars represent standard errors.

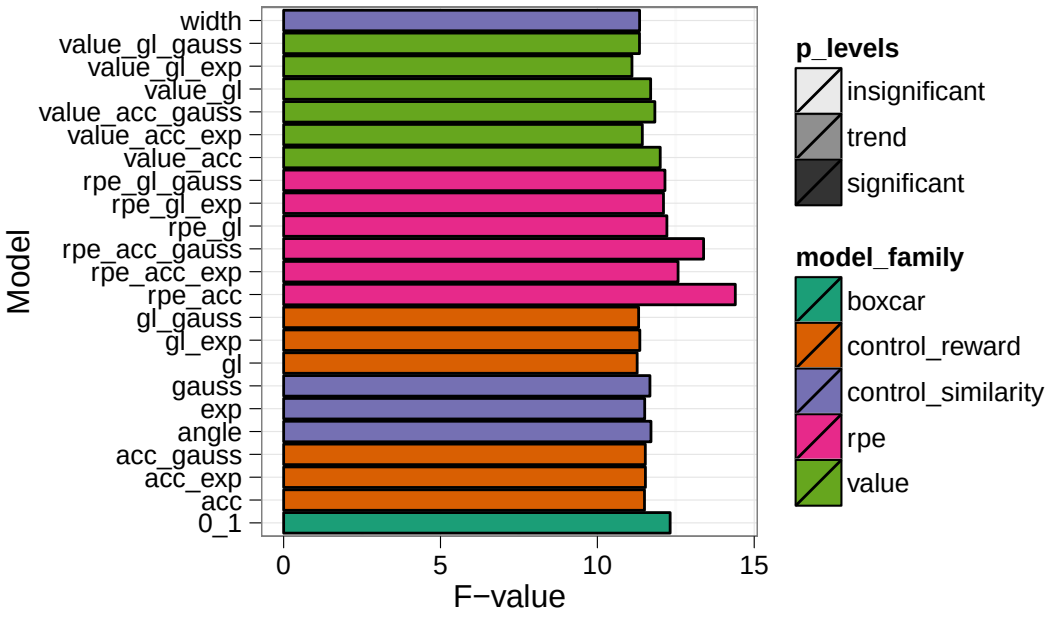

Figure 36. Middle frontal (dorsolateral) PFC – F-values for all models. Significance-level is denoted by the saturation, where the $p < 0.05$ level is significant, and trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p64 for details).

Discussion

The question is, are cognitive rewards represented as categories in the human brain? And does such a representation impact the reinforcement learning process? To start to answer these two interrelated questions, fMRI data was collected while participants completed a stimulus-response task using with pre-trained perceptual categories as rewards, one category for gains and one for losses. Each trial's reward, then, was a never before or again experienced examplar from one of the two reward categories. The use of only new exemplars distinguishes this task from higher-order conditioning paradigms were the same stimulus is repeatedly paired or presented. The behavioral and neural findings of this work, which are now discussed in detail, show that cognitive rewards can be categories, categories which do substantively impact reinforcement learning signals in the brain. It will be further argued that category representations would be a reasonable mechanistic explanation for the generalization of (classical) secondary reinforcers, leading to ultimate conclusion: rewards are categories. To build this case, this discussion will now step through both confirmatory and inconsistent results, beginning with the behavioral and ending with select regions of interest.

Taking Us to Can

In the behavioral task rewards were drawn from perceptual categories, information integration (II) categories to be specific (p29). II is classic category structure, much studied in humans and other animals (J. D. Smith et al., 2011; Ashby & Mad-

dox, 2011; J. D. Smith, Beran, Crossley, Boomer, & Ashby, 2010). II categories are distinct from their contemporaries by requiring integration of multi-dimensional stimulus information, and so are difficult to verbally describe. II learning also recruits procedural memory, which relies heavily on the dorsal striatum (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). The lack of verbalizeability and the multi-dimensional structure make the reward categories irreconcilable with the classical rewards almost universally used human studies of reward (e.g "Win $1", "Correct!", or "Yes!"). Despite the marked difference between classical and the reward categories employed here, participants easily and rapidly learned using the II categories. Performance measures, both accuracy and reaction times, were quite similar to tasks using classical rewards (p32; for comparison see, O'Doherty et al. (2003); Ramnani et al. (2000); Aron et al. (2004); Seger et al. (2010); Seger and Cincotta (2005)).

Further arguing for homology between the reward kinds, the overall pattern of BOLD activity, i.e., all trials compared to the rest trials (p54), was also markedly similar to that observed in nearly identical tasks using classical rewards (for several examples see, Lopez-Paniagua and Seger (2011); Seger et al. (2010); Cincotta and Seger (2007); Seger and Cincotta (2006, 2005)).

Taking in account both the behavioral and neurological consistency observed between classical rewards and the reward categories provides strong initial evidence that previously unexperienced exemplars, from studied categories, *can* acts are rewards. Reversing that logic, rewards therefore *can* be categories. Having established that the next logical question is, do the same neural algorithm(s) that mediate classical reward learning facilitate reward category learning as well?. That is, as reviewed

above (starting on p7), the established mechanistic role for classically rewarding stimuli to drive stimulus-response learning is the reward prediction error hypothesis of phasic dopaminergic firing. So do reward categories effect reward prediction error signals? To ask this question, three kinds of rewards prediction error models were compared (p38). One treated reward exemplars as classical rewards, the other two diminished the reward's worth, and thus the reward prediction error, based on how far that exemplar was from its category mean (i.e., similarity, see p39 for mathematical details).

Are, Reflected in Error(s)

A known pair's logic. However before drawing any conclusions from the modeling data, there are some logical preliminaries to get out of the way. Many of the models of interest are both covariate and dependent. Under generic statistical circumstances it would be difficult, or even impossible, to compare such models. However in limited cases strong, even causal, conclusions are possible. Inside the same family and coding scheme, there is a single change between many of the models. For example, "rpe_acc" and "rpe_acc_gauss" differ only by the similarity adjustment of the reward (i.e., Eq 6 and 9). Because both models are fit to the same data[15] and so have identical signal-to-noise ratios, the 1.5[16] fold increase in information that comes from using "rpe_acc_gauss" in the dorsal caudate *must* be caused by that single change (Pearl, 2010). So while 1.5 would be small increase when comparing two noisy random variables (Anderson et al., 2000; Forster, 2000), it is argued

[15]Using the same deterministic loss function
[16]Bilateral average

that, (1) because uncertainty is constant between the fits, and (2) because we also know the exact relation between two models, and (3) as the model's predictions only sometimes diverge (compare columns in Figure 10), 1.5 should instead be considered strong evidence.

Categories, in all the right spots. In most of the regions of interest, the reward prediction family ("rpe") was the most informative, ranging from 2.3-5.1 times more likely than the non-parametric "boxcar" model (p69). This alone strongly suggests that like classical rewards, the learning driven by reward categories is mediated by the dopaminergic reward prediction signal. Even more important is the fact that many of the most reward sensitive areas are best described by the Gaussian-similarity adjusted reward ("rpe_acc_gauss" in Figures 19, 27, 29, 25, and 31), demonstrating that category parameters (i.e., the similarity metrics) directly affect reward valuation. This is a direct confirmation of the hypothesis that cognitive rewards have an underlying category representation.

Outside the of VTA/SNc, striatal BOLD activity has been, time after time, shown to reflect the dopaminergic reward prediction error signal making it a, if not the, key test of novel reward prediction hypotheses (see the *Introduction* for much supporting evidence on this point). The fact that in the dorsal caudate the Gaussian-adjusted reward prediction error term offered a substantively more informative account than the unadjusted models is a is crucially important result (compare "rpe_acc_gauss" to "rpe_acc" in Figure 19), combined that is with the fact that the dorsal caudate was strongly active (Figure 20) and best described by the "rpe" family

(Figure 19).

The ventral striatum was not well described by any of the models, nor was it significantly active bilaterally (Figure 23 and 24). This is a concern as the ventral striatum was expected to play a strong role in this task, as it is both the ventral and dorsal striatum that have been most often correlated with reward prediction activity (O'Doherty et al., 2003; Knutson & Wimmer, 2007; Schönberg, Daw, Joel, & O'Doherty, 2007). This is not to say that the dorsal and ventral areas are functionally homogeneous (Schonberg et al., 2009; O'Doherty et al., 2004; Atallah, Lopez-Paniagua, Rudy, & O'Reilly, 2007). The dorsal caudate has been repeatedly linked to more abstract kinds of rewarding activity (e.g., task outcomes, fictive rewards; Tricomi and Fiez (2008); Lohrenz et al. (2007); for a review see, Grahn, Parkinson, and Owen (2008)). While ventral activity is often associated with primary rewards, or other hedonic valuations (O'Doherty et al., 2004). Given this functional divide, and the dorsal caudate's established role in II category learning (Ashby et al., 1998), in hindsight perhaps then it is no surprise that only dorsal striatum was found to be active.

The dorsal striatum and ACC have several telling similarities. Both, in part due to dopaminergic projections from the VTA/SNc that modulate LTP via D1 receptors (Schweimer & Hauber, 2006), are strongly involved in cognitive reward learning (Atlas et al., 2010; Hayden et al., 2009; Rudebeck et al., 2008; Rolls, McCabe, & Redoute, 2008; Quilodran, Rothé, & Procyk, 2008; Hampton & O'Doherty, 2007; Ernst et al., 2004), with the BOLD signal often reflecting prediction errors in higher-order conditioning experiments (Seymour et al., 2004) and fictive rewards (Hayden

et al., 2009). The ACC though appears to specialize in mediating between competing *future* alternatives, especially in the context of effort required to achieve each option (Quilodran et al., 2008). The fact then the ACC also is most informatively described by the Gaussian-adjusted reward prediction error is another strong piece of evidence supporting reward category representations.

While generally consistent with the reward category interpretation, the insula was the one region that was equally well described by both reward codes (i.e., "acc": $\{1, 0\}$ or "gl": $\{1, -1\}$). All others strongly preferred "acc". While the functional role of both codes, which are quite different in their predictions (compare Figure 10 to 11, see also Figure **??**), is obscure the "gl" coding is consistent with insula's established role in the processing and prediction of aversive outcomes (Chua, Krams, Toni, Passingham, & Dolan, 1999; Phillips et al., 1998; Büchel, Morris, Dolan, & Friston, 1998; Elliott, Friston, & Dolan, 2000). Additionally, the finding of dual codes in the insula is the first confirmation of the secondary hypothesis (p39), the reported complex reward codes found in single cell recordings of VTA/SNc will be present in the BOLD signal (H. Kim et al., 2006; Matsumoto & Hikosaka, 2009; K. S. Smith et al., 2011).

As reviewed in the *Introduction*, the middle frontal (i.e., dorsolateral) cortex plays role a in estimating future reward probabilities, the singular relation between activity in this region and the "rpe_acc" model (p71, see also Figure 35) is best explained by another of this region's well established roles, the encoding of abstract rules (Wallis, Anderson, & Miller, 2001). While prefrontal regions have been previously shown to reflect prediction errors (Ramnani, Elliott, Athwal, & Passingham,

2004), it is speculates that the reward categories are transformed in dorsolateral PFC into reward rules, something akin to "this category of gratings is worth $1". And that these rule-encoded rewards have their own (separate) reward prediction error calculations.

A fit inconsistency. Both "rpe_acc" and "rpe_gl" fit the behavioral data better either of the corresponding similarity-adjusted models (Figure 7). If rewards are in fact categories the opposite pattern would be expected. This inconsistency though has a strong alternative explanation. Even with perfect performance, the largest possible value estimate is smaller for the adjusted models compared to the unadjusted (as suggested by Figure 15, compare the maximum value peaks for "rpe_acc" compared to "rpe_acc_exp" and "rpe_acc_exp"). These smaller value estimates result in lower probability estimates (via the softmax transform, Eq 11) and thus in lower log-likelihood scores (i.e., worse fits). Despite this inherent limitation the adjusted models could be modified to give equivalent performance. As the task is deterministic, once the optimal choices were learned the models could switch strategies and rely on a "working memory" strategy: just do what you did last time. This kind of working memory has recently been shown to be quite entangled with human reinforcement learning (Collins & Frank, 2012). Alternatively the reward prediction errors could be renormalized based on the cumulative variance, following observations of just such behavior (Tobler et al., 2005).

Back to the secondary. When conditioned as secondary reinforcers simple stimuli generalize well, both in humans and in other animals (for a review see p23).

This generalization is by inference, i.e no direct reinforcement is needed (Guttman, 1956; Nakamura et al., 2006; J. D. Smith et al., 2011). Mechanistically how such generalization occurs has not been studied. Based on the success of the similarity-adjusted reward prediction errors above, it is speculated that the even simple stimuli have fundamentally categorical representations, and that these representations, via similarity-adjusted prediction errors, facilitate stimulus generalization. In addition to perfectly matching Shepard's (1987) theoretical predictions of exponential or Gaussian decays in the degree of generalization (p24), categorical representations, of the kind studied here (p38), for secondary rewards would implicitly allow animals to generalize on the first new example, matching the observed behavior. A categorical basis for even simple stimuli is advantageous in non-generalization trials as well. The intrinsic noise in neuronal coding causes the second viewing a stimulus to have a (slightly) different representation than the first (Ashby & Townsend, 1986). A categorical representation would easily overcome such noisy encodings.

The Big Conclusion

Based on the consistency between classical and reward categories, in both behavioral and overall BOLD activity patterns, it was first concluded that rewards *can* be categories. This, combined with the fact that reward categories generate reward predictions errors, and these errors strongly reflect category structure, and that categories offer a powerful and parsimonious explanation for the generalization of secondary reinforcers, finally concluding that rewards *are* categories.

Future Work

If rewards are categories, the next question is whether rewards are *only* categories. While generalization (and so categories) may be universal (Shepard, 1987), specifics matter too. In fact memory for specifics is at odds with generalizable (i.e., abstract) memories (Atallah, Frank, & O'Reilly, 2004). Given that item and categories must always diverge, rewards with both item and category representations would be useful. However there is a marked degree of overlap between the reward processing and category learning systems (Seger & Miller, 2010; Ashby & Maddox, 2011). While this overlap might be due to the fact that most categories (in the lab) are learned using rewards, there is an alternative, if rewards are just a kind of category. The overlapping activity could reflect only the process of one category building another (dissimilar) category, though see the discussion of dorsolateral PFC above for a first counter to this category-only hypothesis.

If it really is the case that rewards are categories, and category similarity metrics affect reward valuations, then the degree of similarity should be a useful parameter in shaping the learning rate. For example, using the same II category structure (Figure 2) and in a between-groups design, one could select gratings from either closer (group 1) or farther (group 2) to the category means. If learning were slower for group 2 compared to 1, this would be a direct casual confirmation that reward categories drive learning.

Just as there are too many outcomes (i.e., rewards) for a human agent to explore them all, making reward categories such a potentially useful tool for humans

and other animals (reviewed on p22), reward categories could be just as useful for robots, or other computational agents, operating in complex environments. However categories are not naively consistent with the theoretical necessities found in Markov state spaces, into which much of reinforcement learning theory is embedded (Sutton & Barto, 1998). Fortunately the need and study of state generalization is quite an active area or research (?, ?), making it both possible and prudent (for both further theoretical and experimental development) to extend these methods to reward representations as well.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(63), 716–723.

Amaro, E., & Barker, G. J. (2006, Apr). Study design in fmri: basic principles. *Brain Cogn*, *60*(3), 220–32.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *The Journal of Wildlife Management*, *64*(4), 912–923.

Aron, A. R., Shohamy, D., Clark, J., Myers, C., Gluck, M. A., & Poldrack, R. A. (2004, Aug). Human midbrain sensitivity to cognitive feedback and uncertainty during classification learning. *J Physiol.*, *92*(2), 1144–52.

Ashburner, J., & Friston, K. (1999, Jan). Nonlinear spatial normalization using basis functions. *HUMAN BRAIN MAPPING*, *7*, 254—266.

Ashburner, J., Neelin, P., Collins, D., Evans, A., & Friston, K. (1997). Incorporating prior knowledge into image registration. *Neuroimage*, *6*(4), 344–352.

Ashby, F. G., & Alfonso-Reese, L. (1995, Jan). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998, Jul). A neuropsychological theory of multiple systems in category learning. *Psychol Rev*, *105*(3), 442–81.

Ashby, F. G., & Ennis, J. (2006). The role of the basal ganglia in category learning. *Psychology of Learning and Motivation*, *46*, 1–36.

Ashby, F. G., & Maddox, W. T. (2005, Jan). Human category learning. *Annual review of psychology*, *56*, 149–78.

Ashby, F. G., & Maddox, W. T. (2011, Apr). Human category learning 2.0. *Ann N Y Acad Sci*, *1224*, 147–61.

Ashby, F. G., & OBrien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Science*, *9*(2), 83–89.

Ashby, F. G., & Townsend, J. T. (1986, Apr). Varieties of perceptual independence. *Psychol Rev*, *93*(2), 154–79.

Atallah, H. E., Frank, M. J., & O'Reilly, R. C. (2004, Nov). Hippocampus, cortex, and basal ganglia: insights from computational models of complementary learning systems. *Neurobiol Learn Mem*, *82*(3), 253–67.

Atallah, H. E., Lopez-Paniagua, D., Rudy, J. W., & O'Reilly, R. C. (2007, Jan). Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nat Neurosci*, *10*(1), 126–31.

Atlas, L. Y., Bolger, N., Lindquist, M. A., & Wager, T. D. (2010, Sep). Brain mediators of predictive cue effects on perceived pain. *J Neurosci*, *30*(39), 12964–77.

Bassett, D. S., & Bullmore, E. (2006). Small-world brain networks. *The Neuroscientist*, *12*(6), 512–523.

Bayer, H., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*, 129–141.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007, Sep). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–21.

Beierholm, U. R., Anen, C., Quartz, S., & Bossaerts, P. (2011, Jul). Separate encoding of model-based and model-free valuations in the human brain. *NeuroImage*.

Berridge, K. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology*, *191*(3), 391–431.

Berridge, K. C., & Robinson, T. E. (2003, Sep). Parsing reward. *Trends Neurosci*, *26*(9), 507–13.

Birn, R. M., Cox, R. W., & Bandettini, P. A. (2002, Jan). Detection versus estimation in event-related fmri: choosing the optimal stimulus timing. *Neuroimage*, *15*(1), 252–64.

Bischoff-Grethe, A., Hazeltine, E., Bergren, L., Ivry, R. B., & Grafton, S. T. (2009, Jan). The influence of feedback valence in associative learning. *Neuroimage*, *44*(1), 243–51.

Blatter, K., & Schultz, W. (2006, Jan). Rewarding properties of visual stimuli. *Experimental Brain Research*, *168*(4), 541–6.

Blough, D. (2001, Jan). The perception of similarity. *Avian visual cognition*.

Bornstein, A. M., & Daw, N. D. (2011, Mar). Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current opinion in neurobiology*.

Botvinick, M., Niv, Y., & Barto, A. (2008, Oct). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*.

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119–126.

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010a, Jan). Distinct tonic and phasic anticipatory activity in lateral habenula and dopamine neurons. *Neuron*.

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010b, Dec). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, *68*(5), 815–34.

Bromberg-Martin, E. S., Matsumoto, M., Hong, S., & Hikosaka, O. (2010, Aug). A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J Physiol.*, *104*(2), 1068–76.

Bromberg-Martin, E. S., Matsumoto, M., & Nakahara, H. (2010, Jan). Multiple timescales

of memory in lateral habenula and dopamine neurons. *Neuron*.

Büchel, C., Morris, J., Dolan, R. J., & Friston, K. J. (1998, May). Brain systems mediating aversive conditioning: an event-related fmri study. *Neuron*, *20*(5), 947–57.

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., et al. (1996, Feb). Statistical methods of estimation and inference for functional mr image analysis. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, *35*(2), 261–77.

Bunzeck, N., & Düzel, E. (2006, Aug). Absolute coding of stimulus novelty in the human substantia nigra/vta. *Neuron*, *51*(3), 369–79.

Burnham, K. P. (2004, Nov). Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods & Research*, *33*(2), 261–304.

Calabresi, P., Picconi, B., Tozzi, A., & Filippo, M. D. (2007, May). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends Neurosci.*, *30*(5), 211–9.

Cannon, C. M., & Palmiter, R. D. (2003, Nov). Reward without dopamine. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *23*(34), 10827–31.

Chamberlin, T. C. (1965, May). The method of multiple working hypotheses: With this method the dangers of parental affection for a favorite theory can be circumvented. *Science*, *148*(3671), 754–9.

Chatera, N., & Vitányib, P. M. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*(3), 346–369.

Chua, P., Krams, M., Toni, I., Passingham, R., & Dolan, R. (1999, Jun). A functional anatomy of anticipatory anxiety. *Neuroimage*, *9*(6 Pt 1), 563–71.

Cincotta, C. M., & Seger, C. A. (2007, Feb). Dissociation between striatal regions while learning to categorize via feedback and via observation. *Journal of cognitive neuro-*

science, *19*(2), 249–65.

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., & Marchal, G. (1995, Jan). Automated multi-modality image registration based on information theory. *Information Processing in Medical Imaging*, 263–274.

Collins, A. G. E., & Frank, M. J. (2012, Apr). How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *Eur J Neurosci*, *35*(7), 1024–35.

Crow, T. J. (1972, Nov). Catecholamine-containing neurones and electrical self-stimulation. 1. a review of some data. *Psychol Med*, *2*(4), 414–21.

Dale, A. M. (1999, Jan). Optimal experimental design for event-related fmri. *Hum Brain Mapp*, *8*(2-3), 109–14.

Daw, N. D., & Courville, A. (2007). The pigeon as particle filter. *NIPS*, *20*.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011, Mar). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–15.

Dayan, P., & Daw, N. D. (2008, Dec). Decision theory, reinforcement learning, and the brain. *Cognitive, affective & behavioral neuroscience*, *8*(4), 429–53.

Dayan, P., & Niv, Y. (2008, Apr). Reinforcement learning: the good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185–96.

Delgado, M. R., Stenger, V. A., & Fiez, J. A. (2004, Sep). Motivation-dependent responses in the human caudate nucleus. *Cereb Cortex*, *14*(9), 1022–30.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006, Jul). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–80.

Diaconescu, A. O., Menon, M., Jensen, J., Kapur, S., & McIntosh, A. R. (2010, Feb).

Dopamine-induced changes in neural network patterns supporting aversive conditioning. *Brain Res*, *1313*, 143–61.

Dommett, E., Coizet, V., Blaha, C., Martindale, J., Lefebvre, V., Walton, N., et al. (2005, Mar). How visual stimuli activate dopaminergic neurons at short latency. *Science*, *307*(5714), 1476.

Elliott, R., Friston, K. J., & Dolan, R. J. (2000, Aug). Dissociable neural responses in human reward systems. *J Neurosci*, *20*(16), 6159–65.

Ennis, D. (1988). Toward a universal law of generalization. *Science*, *242*(4880), 944.

Ernst, M., Nelson, E. E., McClure, E. B., Monk, C. S., Munson, S., Eshel, N., et al. (2004, Jan). Choice selection and reward anticipation: an fmri study. *Neuropsychologia*, *42*(12), 1585–97.

Fiore, V., Mannella, F., Mirolli, M., Gurney, K., Baldassarre, G., & Ricerche, C. delle. (2008). Instrumental conditioning driven by apparently neutral stimuli: A model tested with a simulated robotic rat. *International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, 139.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003, Mar). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898–902.

Forster, M. (2000, Mar). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology*, *44*(1), 205–231.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001, Jun). Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive, affective & behavioral neuroscience*, *1*(2), 137–60.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004, Dec). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–3.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., & Turner, R. (1998,

Jan). Event-related fmri: characterizing differential responses. *Neuroimage*, *7*(1), 30–40.

Furuyashiki, T., Holland, P. C., & Gallagher, M. (2008, May). Rat orbitofrontal cortex separately encodes response and outcome information during performance of goal-directed behavior. *J Neurosci*, *28*(19), 5127–38.

Geman, S., Bienenstock, E., & Doursat, R. (1, Jan). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Gershman, S., Pesaran, B., & Daw, N. D. (2009, Oct). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *Journal of Neuroscience*, *29*(43), 13524.

Glimcher, P. W., Dorris, M., & Bayer, H. (2005, Aug). Physiological utility theory and the neuroeconomics of choice. *Games and economic behavior*, *52*(2), 213–256.

Grahn, J. A., Parkinson, J. A., & Owen, A. M. (2008, Nov). The cognitive functions of the caudate nucleus. *Progress in Neurobiology*, *86*(3), 141–55.

Guitart-Masip, M., Bunzeck, N., Stephan, K., Dolan, R. J., & Duzel, E. (2010a). Contextual novelty changes reward representations in the striatum. *Journal of Neuroscience*, *30*(5), 1721.

Guitart-Masip, M., Bunzeck, N., Stephan, K., Dolan, R. J., & Duzel, E. (2010b, Feb). Contextual novelty changes reward representations in the striatum. *Journal of Neuroscience*, *30*(5), 1721.

Guttman, N. (1956, Jan). Discriminability and stimulus generalization. *Journal of Experimental Psychology*.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006, Aug). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J Neurosci*, *26*(32), 8360–7.

Hampton, A. N., & O'Doherty, J. P. (2007, Jan). Decoding the neural substrates of reward-related decision making with functional mri. *PNAS*, *104*(4), 1377–82.

Haruno, M., & Kawato, M. (2006, Feb). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *J Physiol.*, *95*(2), 948–59.

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2009, May). Fictive reward signals in the anterior cingulate cortex. *Science*, *324*(5929), 948–50.

Hikosaka, O., Sesack, S. R., Lecourtier, L., & Shepard, P. D. (2008, Nov). Habenula: crossroad between the basal ganglia and the limbic system. *J Neurosci*, *28*(46), 11825–9.

Hnasko, T. S., Sotak, B. N., & Palmiter, R. D. (2005, Dec). Morphine reward in dopamine-deficient mice. *Nature*, *438*(7069), 854–7.

Hollerman, J., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*(4), 304–309.

Hornak, J., O'Doherty, J. P., Bramham, J., Rolls, E. T., Morris, R. G., Bullock, P. R., et al. (2004, Apr). Reward-related reversal learning after surgical excisions in orbito-frontal or dorsolateral prefrontal cortex in humans. *Journal of cognitive neuroscience*, *16*(3), 463–78.

Ito, M., & Doya, K. (2011, Jun). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current opinion in neurobiology*, *21*(3), 368–73.

Iversen, S. D., & Iversen, L. L. (2007). Dopamine: 50 years in perspective. *Trends Neurosci.*, *30*(5), 188–193.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513–541.

Jakel, F., Scholkopf, B., & Wichmann, F. A. (2008, Apr). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, *15*(2), 256–271.

Jay, T. M. (2003). Dopamine: a potential substrate for synaptic plasticity and memory mechanisms. *Progress in Neurobiology*, *69*, 375–390.

Jimura, K., Locke, H. S., & Braver, T. S. (2010, May). Prefrontal cortex mediation of cognitive enhancement in rewarding motivational contexts. *Proc Natl Acad Sci USA*, *107*(19), 8871–6.

Jin, X., & Costa, R. M. (2010, Jul). Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature*, *466*(7305), 457–462.

Joel, D., Niv, Y., & Ruppin, E. (2002, Jan). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, *15*(4-6), 535–47.

Kahnt, T., Heinzle, J., Park, S. Q., & Haynes, J.-D. (2010, May). Decoding different roles for vmpfc and dlpfc in multi-attribute decision making. *Neuroimage*.

Kakade, S., & Dayan, P. (2002, Jan). Dopamine: generalization and bonuses. *Neural Networks*, *15*(4-6), 549–559.

Kao, M.-H., Mandal, A., Lazar, N., & Stufken, J. (2009, Feb). Multi-objective optimal experimental designs for event-related fmri studies. *Neuroimage*, *44*(3), 849–56.

Kim, H., Shimojo, S., & O'Doherty, J. P. (2006, Jul). Is avoiding an aversive outcome rewarding? neural substrates of avoidance learning in the human brain. *PLoS Biology*, *4*(8), e233.

Kim, H., Shimojo, S., & O'Doherty, J. P. (2010, Aug). Overlapping responses for the expectation of juice and money rewards in human ventromedial prefrontal cortex. *Cereb Cortex*.

Kim, S., Hwang, J., & Lee, D. (2008, Jul). Prefrontal coding of temporally discounted

values during intertemporal choice. *Neuron*, *59*(1), 161–72.

Kim, S., Hwang, J., Seo, H., & Lee, D. (2009, Apr). Valuation of uncertain and delayed rewards in primate prefrontal cortex. *Neural Networks*, *22*(3), 294–304.

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001, Aug). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J Neurosci*, *21*(16), RC159.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005, May). Distributed neural representation of expected value. *J Neurosci*, *25*(19), 4806–4812.

Knutson, B., & Wimmer, G. E. (2007, May). Splitting the difference: how does the brain code reward episodes? *Annals of the New York Academy of Sciences*, *1104*, 54–69.

Krebs, R. M., Heipertz, D., Schuetze, H., & Duzel, E. (2011, Jun). Novelty increases the mesolimbic functional connectivity of the substantia nigra/ventral tegmental area (sn/vta) during reward anticipation: Evidence from high-resolution fmri. *NeuroImage*.

Kruggel, F., Cramon, D. Y. V., & Descombes, X. (1999). Comparison of filtering methods for fmri datasets. *Neuroimage*, *10*(5), 530–543.

Kruschke, J. (n.d.). Alcove: An exemplar-based connectionist model of category learning. *PSYCHOLOGICAL REVIEW*, *99*(1), 22–44.

Lazarus, M., Shen, H.-Y., Cherasse, Y., Qu, W.-M., Huang, Z.-L., Bass, C. E., et al. (2011, Jul). Arousal effect of caffeine depends on adenosine a2a receptors in the shell of the nucleus accumbens. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *31*(27), 10067–10075.

Lisman, J. E., & Grace, A. A. (2005, Jun). The hippocampal-vta loop: controlling the entry of information into long-term memory. *Neuron*, *46*(5), 703–13.

Liu, T. T. (2004, Jan). Efficiency, power, and entropy in event-related fmri with multiple

trial types. part ii: design of experiments. *Neuroimage*, *21*(1), 401–13.

Lohrenz, T., McCabe, K., Camerer, C., & Montague, P. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences*, *104*(22), 9493.

Lopez-Paniagua, D., & Seger, C. A. (2011). Interactions within and between corticostriatal loops during component processes of category learning. *Journal of cognitive neuroscience*, *23*(10), 3068–3083.

Love, B. C. (2004, Jan). Sustain: A network model of category learning. *PSYCHOLOGICAL REVIEW*, 309—332.

Maddox, W. T., & Bohil, C. J. (2001, Jun). Feedback effects on cost-benefit learning in perceptual categorization. *Mem Cognit*, *29*(4), 598–615.

Mars, R. B., Shea, N. J., Kolling, N., & Rushworth, M. F. S. (2010, Apr). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Q J Exp Psychol (Colchester)*, 1–16.

Martin, J. B., Griffiths, T. L., & Sanborn, A. N. (2012, Jan). Testing the efficiency of markov chain monte carlo with people using facial affect categories. *Cogn Sci*, *36*(1), 150–62.

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, *459*(7248), 837–841.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003, Apr). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–46.

McKinley, S. C., & Nosofsky, R. M. (1996, Apr). Selective attention and the formation of linear decision boundaries. *J Exp Psychol Hum Percept Perform*, *22*(2), 294–317.

Medin, D., & Schaffer, M. (n.d.). Context theory of classification learning. *Psychological Review*, *85*(3), 207.

Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., & Buckner, R. L. (2000, Jun). Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *Neuroimage*, *11*(6 Pt 1), 735–59.

Mirenowicz, J., & Schultz, W. (1994). Importance of unpredictability for reward responses in primate dopamine neurons. *J Physiol.*, *72*(2), 1024.

Mishra, A. M., Ellens, D. J., Schridde, U., Motelow, J. E., Purcaro, M. J., DeSalvo, M. N., et al. (2011, Oct). Where fmri and electrophysiology agree to disagree: corticothalamic and striatal activity patterns in the wag/rij rat. *J Neurosci*, *31*(42), 15053–64.

Montague, P., King-Casas, B., & Cohen, J. (2006). Imaging valuation models in human choice. *Annu Rev Neurosci*, *29*, 417–448.

Nakamura, T., Ito, M., Croft, D. B., & Westbrook, R. F. (2006, Nov). Domestic pigeons (columba livia) discriminate between photographs of male and female pigeons. *Learning & Behavior*, *34*(4), 327–339.

Nosofsky, R. (1985, Nov). Overall similarity and the identification of separable-dimension stimuli: a choice model analysis. *Percept Psychophys*, *38*(5), 415–32.

Nosofsky, R. (1988, Oct). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708.

Nosofsky, R. M. (1988, Jan). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 54—65.

O'Doherty, J. P. (2003, Aug). Can't learn without you: predictive value coding in orbitofrontal cortex requires the basolateral amygdala. *Neuron*, *39*(5), 731–3.

O'Doherty, J. P., Buchanan, T. W., Seymour, B., & Dolan, R. J. (2006, Jan). Predictive

neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. *Neuron*, *49*(1), 157–66.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003, Apr). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–37.

O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004, Apr). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–4.

O'Doherty, J. P., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001, Jan). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat Neurosci*, *4*(1), 95–102.

O'Reilly, R., Frank, M. J., Hazy, T., & Watz, B. (2007, Jan). Pvlv: the primary value and learned value pavlovian learning algorithm. *Behav. Neurosci*.

Ouden, H. den, Daunizeau, J., Roiser, J., Friston, K., & Stephan, K. (2010). Striatal prediction error modulates cortical coupling. *Journal of Neuroscience*, *30*(9), 3210.

Packard, M. G., & Knowlton, B. J. (2002, Jan). Learning and memory functions of the basal ganglia. *Annu Rev Neurosci*, *25*, 563–93.

Pearl, J. (2010, Jan). An introduction to causal inference. *Int J Biostat*, *6*(2), Article 7.

Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, V., et al. (1998, Oct). Neural responses to facial and vocal expressions of fear and disgust. *Proc Biol Sci*, *265*(1408), 1809–17.

Pizzagalli, D. A., Evins, A. E., Schetter, E. C., Frank, M. J., Pajtas, P. E., Santesso, D. L., et al. (n.d.). Single dose of a dopamine agonist impairs reinforcement learning in humans: Behavioral evidence from a laboratory-based measure of reward responsiveness. *Psychopharmacology*, *196*(2), 221.

Pizzagalli, D. A., Evins, A. E., Schetter, E. C., Frank, M. J., Pajtas, P. E., Santesso, D. L., et al. (2008, Feb). Single dose of a dopamine agonist impairs reinforcement learning in humans: Behavioral evidence from a laboratory-based measure of reward responsiveness. *Psychopharmacology (Berl)*, *196*(2), 221–232.

Poldrack, R. A. (2007, Mar). Region of interest analysis for fmri. *Social Cognitive and Affective Neuroscience*, *2*(1), 67–70.

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008, Apr). Guidelines for reporting an fmri study. *Neuroimage*, *40*(2), 409–14.

Poldrack, R. A., & Foerde, K. (2008, Jan). Category learning and the memory systems debate. *Neuroscience and Biobehavioral Reviews*, *32*(2), 197–205.

Quilodran, R., Rothé, M., & Procyk, E. (2008, Jan). Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron*, *57*(2), 314–25.

Ramnani, N., Elliott, R., Athwal, B. S., & Passingham, R. E. (2004, Nov). Prediction error for free monetary reward in the human prefrontal cortex. *Neuroimage*, *23*(3), 777–86.

Ramnani, N., Toni, I., Josephs, O., Ashburner, J., & Passingham, R. E. (2000, Dec). Learning- and expectation-related changes in the human brain during motor learning. *J Physiol.*, *84*(6), 3026–35.

Rao, C. R., Wu, Y., Konishi, S., & Mukerjee, R. (2001). On model selection. *Lecture Notes-Monograph Series, Model Selection*, *38*, 1–64.

Redish, A. D. (2004, Dec). Addiction as a computational process gone awry. *Science*, *306*(5703), 1944–7.

Reed, P. (1992). Signalled delay of reward: Overshadowing versus sign- tracking explanations. *Learning and Motivation*, *23*(1), 27–42.

Rescorla, R. A. (1988, Mar). Pavlovian conditioning. it's not what you think it is. *Am*

Psychol, *43*(3), 151–60.

Robinson, T. E., & Berridge, K. C. (1993, Jan). The neural basis of drug craving: an incentive-sensitization theory of addiction. *Brain Res Brain Res Rev*, *18*(3), 247–91.

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007, Dec). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat Neurosci*, *10*(12), 1615–24.

Roesch, M. R., & Olson, C. R. (2007, Dec). Neuronal activity related to anticipated reward in frontal cortex: does it represent value or reflect motivation? *Annals of the New York Academy of Sciences*, *1121*, 431–46.

Rolls, E. T., McCabe, C., & Redoute, J. (2008, Mar). Expected value, reward outcome, and temporal difference error representations in a probabilistic decision task. *Cereb Cortex*, *18*(3), 652–63.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350.

Rossee, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*(2), 178–210.

Rossi, S., Chiara, V. D., Musella, A., Mataluni, G., Sacchetti, L., Siracusano, A., et al. (2010). Effects of caffeine on striatal neurotransmission: Focus on cannabinoid cb1 receptors. *Molecular Nutrition & Food Research*, *54*(4), 525–531.

Rudebeck, P. H., Behrens, T. E., Kennerley, S. W., Baxter, M. G., Buckley, M. J., Walton, M. E., et al. (2008, Dec). Frontal cortex subregions play distinct roles in choices between actions and stimuli. *J Neurosci*, *28*(51), 13775–85.

Salamone, J. D., Correa, M., Mingote, S. M., & Weber, S. M. (2005, Feb). Beyond the reward hypothesis: alternative functions of nucleus accumbens dopamine. *Curr Opin Pharmacol*, *5*(1), 34–41.

Schmitzer-Torbert, N., & Redish, A. D. (2004, May). Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the multiple t task. *J Physiol.*, *91*(5), 2259–72.

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007, Nov). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J Neurosci*, *27*(47), 12860–7.

Schonberg, T., O'Doherty, J. P., Joel, D., Inzelberg, R., Segev, Y., & Daw, N. D. (2009, Aug). Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in parkinson's disease patients: evidence from a model-based fmri study. *Neuroimage*.

Schultz, W. (2007). Behavioral dopamine signals. *Trends Neurosci.*, *30*(5), 203–210.

Schweimer, J., & Hauber, W. (2006, Jan). Dopamine d1 receptors in the anterior cingulate cortex regulate effort-based decision making. *Learn Mem*, *13*(6), 777–82.

Seger, C. A. (2008, Jan). How do the basal ganglia contribute to categorization? their roles in generalization, response selection, and learning via feedback. *Neuroscience and Biobehavioral Reviews*, *32*(2), 265–78.

Seger, C. A., & Cincotta, C. (2005). The roles of the caudate nucleus in human classification learning. *J Neurosci*, *25*(11), 2941–2951.

Seger, C. A., & Cincotta, C. M. (2006, Nov). Dynamics of frontal, striatal, and hippocampal systems during rule learning. *Cereb Cortex*, *16*(11), 1546–55.

Seger, C. A., & Miller, E. K. (2010, Jan). Category learning in the brain. *Annu Rev Neurosci*, *33*, 203–19.

Seger, C. A., Peterson, E. J., Cincotta, C. M., Lopez-Paniagua, D., & Anderson, C. W. (2010, Apr). Dissociating the contributions of independent corticostriatal systems to visual categorization learning through the use of reinforcement learning modeling

and granger causality modeling. *Neuroimage*, *50*(2), 644–56.

Seymour, B., Daw, N. D., Dayan, P., Singer, T., & Dolan, R. J. (2007, May). Differential encoding of losses and gains in the human striatum. *Journal of Neuroscience*, *27*(18), 4826.

Seymour, B., O'Doherty, J. P., Dyan, P., Koltzenburg, M., Jones, J. K., Dolan, R. J., et al. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*, 664 – 667.

Shepard, R. (1987, Sep). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Simmons, S. (2008, Jan). Individual differences in the perception of similarity and difference. *Cognition*.

Smith, A., Li, M., Becker, S., & Kapur, S. (2006, Mar). Dopamine, prediction error and associative learning: a model-based account. *Network*, *17*(1), 61–84.

Smith, E. E., & Grossman, M. (2008, Jan). Multiple systems of category learning. *Neuroscience and Biobehavioral Reviews*, *32*(2), 249–64.

Smith, J. D., Ashby, F. G., Berg, M. E., Murphy, M. S., Spiering, B., Cook, R. G., et al. (2011). Pigeons' categorization may be exclusively nonanalytic. *Psychonomic Belletin & Review*, *18*(2), 414–421.

Smith, J. D., Beran, M. J., Crossley, M. J., Boomer, J., & Ashby, F. G. (2010, Jan). Implicit and explicit category learning by macaques (macaca mulatta) and humans (homo sapiens). *J Exp Psychol Anim Behav Process*, *36*(1), 54–65.

Smith, K. S., Berridge, K. C., & Aldridge, J. W. (2011, Jul). Disentangling pleasure from incentive salience and learning signals in brain reward circuitry. *Proc Natl Acad Sci USA*, *108*(27), E255–64.

Spanagel, R., & Weiss, F. (1999, Nov). The dopamine hypothesis of reward: past and

current status. *Trends Neurosci.*, *22*(11), 521–7.

Spiering, B. J., & Ashby, F. G. (2008, Sep). Response processes in information-integration category learning. *Neurobiol Learn Mem*, *90*(2), 330–8.

Surmeier, D. J., Ding, J., Day, M., Wang, Z., & Shen, W. (2007, May). D1 and d2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends Neurosci.*, *30*(5), 228–35.

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

Tanaka, S. C., Balleine, B. W., & O'Doherty, J. P. (2008, Jun). Calculating consequences: brain systems that encode the causal effects of actions. *J Neurosci*, *28*(26), 6750–5.

Tanaka, S. C., Samejima, K., Okada, G., Ueda, K., Okamoto, Y., Yamawaki, S., et al. (2006). Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics. *Neural Networks*, *19*(8), 1233–1241.

Tobler, P. N., Christopoulos, G. I., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2009, Apr). Risk-dependent reward value signal in human prefrontal cortex. *Proceedings of the National Academy of Sciences*, *106*(17), 7185–7190.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005, Mar). Adaptive coding of reward value by dopamine neurons. *Science*, *307*(5715), 1642–5.

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2007, Feb). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *J Physiol.*, *97*(2), 1621–32.

Tricomi, E., & Fiez, J. A. (2008, Jul). Feedback signals in the caudate reflect goal achievement on a declarative memory task. *Neuroimage*, *41*(3), 1154–67.

Urcuioli, P. (2001, Jan). Categorization and acquired equivalence. *Avian visual cognition [On-line]. Available: www. ....*

Waelti, P., Dickinson, A., & Schultz, W. (2001, Jul). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*(6842), 43–8.

Wagenmakers, E.-J., & Farrell, S. (2004, Feb). Aic model selection using akaike weights. *Psychon Bull Rev*, *11*(1), 192–6.

Wager, T. D., & Nichols, T. E. (2003, Feb). Optimization of experimental design in fmri: a general framework using a genetic algorithm. *Neuroimage*, *18*(2), 293–309.

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001, Jun). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*(6840), 953–6.

Wimmer, G. E., Daw, N. D., & Shohamy, D. (2012, Apr). Generalization of value in reinforcement learning by humans. *Eur J Neurosci*, *35*(7), 1092–104.

Winterbauer, N. E., & Balleine, B. W. (2007, Jan). The influence of amphetamine on sensory and conditioned reinforcement: evidence for the re-selection hypothesis of dopamine function. *Frontiers in integrative neuroscience*, *1*, 9.

Wise, R. A. (1978, Aug). Catecholamine theories of reward: a critical review. *Brain Res*, *152*(2), 215–47.

Wittmann, B. C., Bunzeck, N., Dolan, R. J., & Düzel, E. (2007, Oct). Anticipation of novelty recruits reward system and hippocampus while promoting recollection. *Neuroimage*, *38*(1), 194–202.

Wittmann, B. C., Daw, N. D., Seymour, B., & Dolan, R. J. (2008, Jun). Striatal activity underlies novelty-based choice in humans. *Neuron*, *58*(6), 967–73.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996, Jan). A unified statistical approach for determining significant signals in images of cerebral activation. *Hum Brain Mapp*, *4*(1), 58–73.

Yin, H. H., & Knowlton, B. J. (2006, Jun). The role of the basal ganglia in habit formation. *Nat Rev Neurosci*, *7*(6), 464–76.

Yin, H. H., Ostlund, S. B., & Balleine, B. W. (2008, Oct). Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *Eur J Neurosci*, *28*(8), 1437–48.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005, Jul). The role of the dorsomedial striatum in instrumental conditioning. *Eur J Neurosci*, *22*(2), 513–23.

Zink, C. F., Pagnoni, G., Chappelow, J., Martin-Skurski, M., & Berns, G. S. (2006, Feb). Human striatal activation reflects degree of stimulus saliency. *Neuroimage*, *29*(3), 977–83.

Zink, C. F., Pagnoni, G., Martin, M. E., Dhamala, M., & Berns, G. S. (2003, Sep). Human striatal response to salient nonrewarding stimuli. *J Neurosci*, *23*(22), 8092–7.

Zink, C. F., Pagnoni, G., Martin-Skurski, M. E., Chappelow, J. C., & Berns, G. S. (2004, May). Human striatal responses to monetary reward depend on saliency. *Neuron*, *42*(3), 509–17.