Categories of rewards.

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

Prologue

If from a dopamenergic point of view novelty *is* a reward instead of being *like* a reward I, and directly others (Kakade & Dayan, 2002), predicted that when positive and novel outcomes combined, they would act sympathetically to accelerate learning compared to either alone. Likewise negative and novel outcomes would interfere, diminishing learning. Using a simple two-choice abstract category learning task, three consecutive experiments were consistent with these predictions leading to the proposal discussed and approved last we met. I am now forced to conclude, after another 20 experiments, that these promising positive results were (at best) highly brittle - in total less the 12% of participants exhibited the effect and I have been unable to reproduce it at all in the last 10 or so experiments, two of which were direct replications of previous successful trials. Analysis of the combined datasets from nearly all 23 experiments suggested the promising three were driven by noise alone, even though each of the three closely approached or exceeded $p < 0.05$ threshold, even when using non-parametric statistics[1]. Having spent a year then on a windmill hunt, I stopped to rethink.

The motivation for studying novelty and reward had two facets. First, much of work with reinforcement learning and dopamine has been correlative, highly correlative in some cases, but still. The exceptions, all prior casual studies in humans have used doses of dopaminergic drugs whose effects are (1) not well understood physiologically and (2) extend beyond the VTA/SNc (i.e. the dopamenergic midbrain) and the striatum into all of cortex (Menon et al., 2007; Pizzagalli et al., 2008; Schonberg et al., 2009). That is while drug-based experiments show casual behavioral effects

[1] Wilcox rank sum, to be specific.

consistent with the reward predictions error (RPE) hypothesis, their specificity is doubtful. By holding everything constant but the presence or absence of novelty the prior proposal allowed for a causal test of the relation between reward, novelty and learning in healthy young adults (as well as a quantitative test via the novelty-extended RPE equations provided by Kakade & Dayan, 2002). Given the noise and reliability problems though this plan must disappear.

The second motivation though was, in a sense, larger. As anything in principle, and so far I am aware in practice, can be novel (i.e. contextually unexpected) but not everything can be linked to primary or secondary reinforcers (i.e. rewards Björklund & Dunnett, 2007), at least not without driving their potency to zero, then if novelty is a reward it must derived from another source. This source is then likely (somewhat) independent of the usual associations to evolutionary primitives such as food, pain and sex. It would instead seem to require evaluation of and derivation from current sense experience and past memories, from an endogenous criterion alone. Reviewing again the literature with cognitive rewards in mind lead to other separate examples of rewards that also seem cognitive (see the introduction for more). In fact, that are more than enough examples to assert cognitive rewards are fact. So then I began to wonder what other under-appreciated or undiscovered complexities reward representations possess. Cognitive rewards could certainly be more flexible; perhaps instead of being coded as just examples, each tasty sip of juice is a wholly independent neural entity, perhaps instead rewards are (or can be at least) categories. That is while rewards can facilitate category learning perhaps, quasi-paradoxically, they are categories themselves; a selective reading of past research is certainly suggestive....

Introduction

This introduction is tripartite. First I make a case for cognitive rewards then and the utility of categorical reward representations. Second is a discussion of prior studies of instrumental generalization (in pigeons) multi-valued decision making (in humans) – both of which can be viewed to a limited degree as evidence for categorical reward representations. And in the final paragraph, I outline the exact methods and goals of this proposed work.

TODO: *Cognitive:* Dicuss, Tricomi (Tricomi & Fiez, 2008), the two fictive reward papers (Hayden, Pearson, & Platt, 2009; Lohrenz, McCabe, Camerer, & Montague, 2007; Dayan, Niv, Seymour, & Daw, 2006), and Dayan's information about outcome report (Bromberg-Martin & Hikosaka, 2009). Also discuss how fictive

rewards and Tricomi's report may suggest categorical rewards - when do subjects really have a specific outcome in mind or a general sense or category of outcomes that would serve?

TODO: *Category-ish*: Devote 2 short paragraphs to instrumental generalization in pigeons and the one multi-valued decision fMRI paper (Kahnt, Heinzle, Park, & Haynes, 2010). Add in the better option firing in rats (Roesch, Calu, & Schoenbaum, 2007)? Make the point at the end that none of these were tested as proper categories (which imply a generalization to never before experienced events), but were instead just previously viewed simple visual information presented in new contexts.

TODO: *Finally*: The task painted as one big picture....also not that we are jittering between stimulus-response and feedback portions of each trial.

Methods and Analyses

Task

As discussed at the end of the introduction, the experimental procedure consists of two parts or tasks. Depicted in Fig (top), the first is a passive classical conditioning where participants will learn reward categories by pairing randomly selected (without replacement) black and white sinusoidal gratings with "Gain $1" or "Lose $1" in, respectively, green or red letters. Reward categories will be derived from a information integration parameter distribution (Fig.). The grating is onscreen for 0.5 seconds, followed by 0.5 seconds gap, with the outcome displayed for another 0.5 seconds with a 1 second fixation cross between each trial. Task 1, which will be completed outside the fMRI scanner, lasts just over 6 minutes displaying 175 examples approximately evenly divided among the two categories. Each participant will be instructed to "Attend to the screen in order to learn which types of gratings lead to winning money and which types lead to losses". The category distribution to value mapping will randomized for each participant.

The second task (Fig , bottom) is an abstract deterministic category learning task that replace direct feedback with the grating from task 1 where each trial begins with appearance of an abstract black and white "tree" stimuli, which belongs to one of two arbitrarily named categories ("q" or "w") of which one if selected by button press using either the right index or middle finger (respectively) on a magnet compatible response box placed on the participants thigh. The response window lasts up to 2.5

seconds. In this period the "tree" is always onscreen. Once either time has elapsed or the participant responds, the "tree" disappears and 1-8 seconds later (as defined by the jitter optimization routine, below) is replaced by a sinusoidal grating. If the response was correct a new, that is never before experienced, exemplar grating from the "gain" distribution is used; if it was incorrect, a new "lose" grating appears. The feedback grating stays onscreen for 1 second and is then immediately replaced by a fixation cross, whose duration is again governed by the jitter optimization routine. Participants will learn to classify 6 "trees" (randomly selected at the start of the experiment out of a pool of 16). Each of the 6 are experienced a total of 40 times for total of 240 trials/scanning session. In this task participants are in part instructed to, "Use what you learned about the rewarding properties of the gratings to try and earn as much money as possible in this portion of the experiment". Instructions for both tasks are verbal using Fig as a visual aid.

In fMRI (and really time-series signal analysis in general) there is an intrinsic tradeoff between simply detecting that a signal has occurred in the presence of noise and then estimating the timecourse (i.e. shape) of that signal (Dale, 1999; Birn, Cox, & Bandettini, 2002; Liu, 2004). The state of the art method for setting trial ordering in an attempt to maximize both signal detection and estimation is a genetic algorithm design by Kao, Mandal, Lazar, & Stufken, 2009. Without extensive modification unfortunately their methodology cannot account for epoch-style designs[2] that this experiment requires; each trial needs to contain stimulus-response, jitter and feedback delivery periods. To account for this, Koa's methods will be applied twice, once using a 6.5 second ISI (the estimated average length of a trial) and once with a 1 second ISI (the length of the feedback display in task 2). These two designs will then be manually interpolated to create a series of stimulus-response, jitter, feedback, and ITI events. To create the final trial ordering each stimulus-response event will then be subsequently randomly recoded to match one of the 6 "tree" stimuli. This randomization step will be redone for each subject.

Finally, as not all participants can successfully learn task 2 (see *Behavioral Results* below) subjects will be prescreened with a variation of task 2, using colored fractals as stimuli, "a" and "b" as category labels, and written feedback (e.g. "Correct"). To be included in the study participants must reach 0.65 accuracy (measured with a rolling average) within 60 training examples during prescreening.

[2]This flaw is common to all methods of stimulus timing optimization, so far as I am aware.

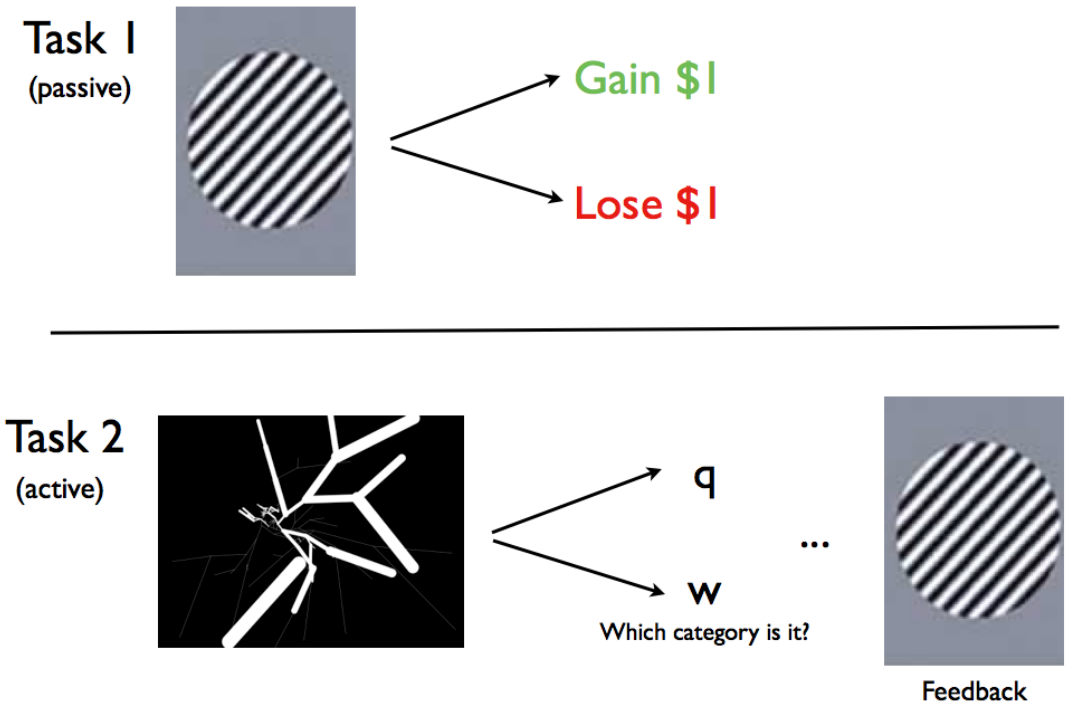

Figure 1. Depiction of the behavioral task. The top is the (passive) classical conditioning participants learn the reward categories. The bottom is the active abstract two-choice category learning.

Figure 2. Sinusoidal grating distributions. **TODO** Describe the parameters of II and why I use it here

fMRI acquisition and analyses

fMRI data for about 15 participants will be acquired at U.C. Boulder (approximately half female, age range 18-35, right handed). All participants will be prescreened for the typical exclusion factors (e.g. metal implants, mental disorders, etc) and will be compensated at a base rate of $15.00 earning up to $30 more depending on behavioral performance. 30 trials will be randomly selected from the 240 possible and the participant will be paid an additional dollar for every correct response and will lose a dollar for every incorrect. The exact scanning parameters (TR, TE, in-plane resolution, etc) are to be determined as this is the Lab's first use of the Boulder machine but will be typical of a whole-brain rapid event-related fMRI

acquisition.

Standard fMRI data preprocessing (motion correction, slice-time correction, drift correction, temporal, spatial smoothing and Tailarach normalization) will be carried out in BrainVoyager, version 2.1. And while the main focus of this work is to test how reward categories are represented and processed, this is so far as I am aware, the first time rewards have been treated as categories in an fMRI experiment, as such basic whole brain activation mapping is also of interest. This too will be done in BrainVoyager with the final maps reported as set of two overlaid probability maps - heat maps where the value in each voxel is the fraction of participants who had significant activity (defined here as either $p < 0.05$ or $p < 001$; RFX correction) at that voxel. This is an improvement over the typical thresholded t-value maps as probability maps include information not just about the strength of the effect but also its consistency.

A priori whole brain contrasts of interest are all trials compared to fixation (the most powerful contrast defining an overall guide to activity patterns), all stimulus-response periods compared to baseline as well as all feedback periods compared to fixation (what were the trial-jittered contributions to the overall signal) as well as feedback contrasted to stimulus-response (and the reverse). Finally the effect of outcome valance (gain of lose) will be compared for each of the prior contrasts. Finally probability maps will be qualitatively compared to results from a similar experiment (using verbal feedback - "correct" or "incorrect") carried out by Lopez-Paniagua, PhD, for his masters work. *A priori* regions of interest are the VTA/SNc, ventral and dorsal striatum, ventromedial and dorsolateral PFC (the latter two have been show to correlate with reinforcement learning value computation Kahnt et al., 2010). All regions will be define by anatomical tracings based of an average of all participants anatomical MRI scans.

Computational analyses

Reinforcement learning measures will be derived from a Rescorla-Wagner models (Eq. 1 and 2.), with decision making approximated by the logistic function (i.e. softmax, Eq. 7). The parameters (α and β) are minimized by maximum log-likelihood. Two separate Rescorla-Wagner models will be considered; is the reward representation that will change (Eq. 3 and 4). We test whether the BOLD signal in the ventral striatum and VTA/SNc is best described by the prediction error term (Eq 2) combined with with Eq. 3 or with Eq. 4. Where D is the Euclidean distance from the average angle ($\bar{\theta}$) and width ($\bar{W}$) in one of the two reward categories ver-

sus a given parameter set for a single example grating (i.e. θ and w). Similarly BOLD changes in the two prefrontal regions of interest (see *fMRI*) will be modeled by $V(s,t)$ (Eq 1), again comparing the two methods for calculating r(t).

To restate, Rescorla-Wagner value updates are defined by,

$$V(s,t) \leftarrow V(s,t) + \alpha * \delta \tag{1}$$

where

$$\delta = r(t) - V(s,t) \tag{2}$$

with $r(t)$ calculated as either

$$r_c(t) = \{1, 0\} \tag{3}$$

or

$$r_d(t) = \frac{r_c(t)}{D} \tag{4}$$

where

$$D = \sqrt{(\bar{\theta} - \theta)^2 + (\bar{W} - w)^2} \tag{5}$$

and in all cases

$$V_{initial}(s,t) = 0. \tag{6}$$

$$p(s_1) = \frac{e^{\beta V(s_1,t)}}{e^{\beta V(s_1,a)} + e^{\beta V(s_2,a)}};\ s_1 = (s_i, q),\ s_2 = (s_i, w). \tag{7}$$

Prior to regression analysis, for each regions of interest outlined above, all reinforcement measures will be convolved with the canonical ("double-gamma") hemodynamic response function and to be consistent with the treatment of BOLD data, z-scored and 4Hz high-pass filtered. Model fits will be compared by AIC and BIC[3]. Assuming AIC and BIC agree (as is likely) the best fit model will only be accepted as valid if it is also significant predictor in the regression.

TODO statistical mediation analysis plan - motivate and briefly describe.

Two quick notes: First, all models in this work assume reward firing in both VTA/SNc and ventral striatum is bivalent, i.e. positive outcomes lead to an increase in firing and negative outcomes lead to a suppression. This is assumption has been

[3]There is so far as I can tell some debate over the best criterion, in general, and it seems there is no harm in calculating both and hoping for agreement. In case of disagreement I propose using their F1 score to decide the best model fit ($F1 := \frac{AIC*BIC}{AIC+BIC}$)

made by nearly all RPE work to date so is not unjustified. However recent recordings in monkeys suggest that there are multiple sub populations. Some bivalent, some firing positively to both both positive and negative, and some inverted, suppressing to positive firing to negative. Even among these subpopulations individual neurons showed trial-by-trial variance. So without further information it would difficult, nye impossible, to model these complex and opposing patterns as a single timecourse model as is necessary for regression analysis. Second, using data exported from BrainVoyager in the NIFTI format (http://nifti.nimh.nih.gov/) all computational analyses will be carried out using custom Python (2.7.1) code developed for this proposal, as well as ongoing unrelated machine learning (MVPA) and fMRI simulation experiments. Code is complete and ready for use, excepting that of statistical mediation testing. Code for this will be ported from the Wagner Lab's related Matlab toolbox (available at http://wagerlab.colorado.edu/tools).

Behavioral results

The experiment protocol outlined above was completed by 33 participants (2/3 female; see Fig and). Using 63% as the cutoff which is near the $p < 0.05$ threshold of the binomial test learning significantly exceeded chance at trial 19. This learning rate is consistent with past work in the lab using just verbal or monetary feedback. Also consistent with prior work was the mean reaction time remaining steady (near 750 ms) as learning proceeded. In summary, the categorical rewards appear behaviorally very similar to direct verbal or monetary rewards.

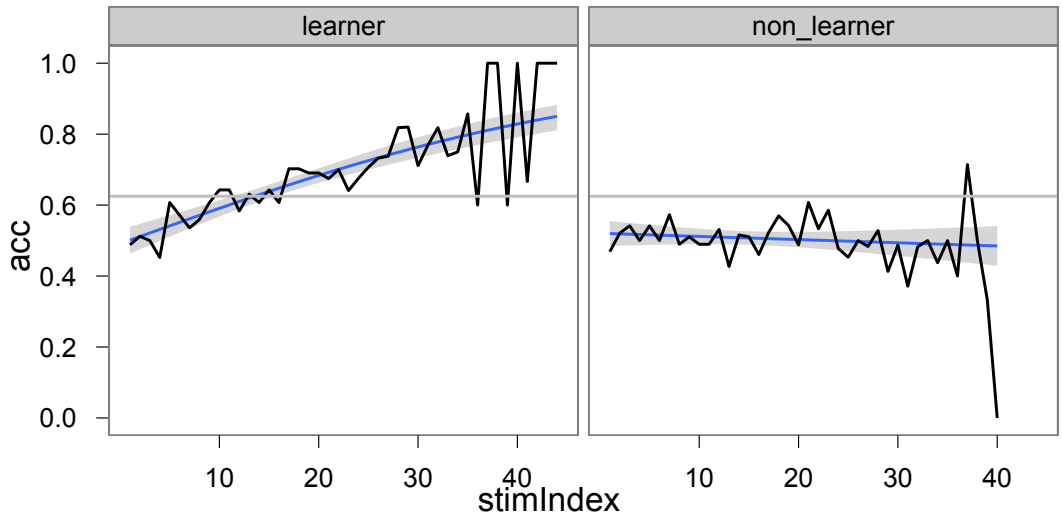

Figure 3. Mean accuracy (black), averaged for all 6 stimuli by trial, blue line and grey represent a binomial regression fit of the data and bootstrapped 95% confidence intervals, respectively. 53% were learners (left) defined by a mean accuracy in the last 10 trials (for all 6 stimuli) greater than 63%. Note: the trial order randomization procedure in this pilot did not enforce equal number of trials in for each of the 6 stimuli (M=38) leading to artifactual increases in variability above trial 35 or so. This oversight will be corrected prior to fMRI data acquisition.

References

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.

Björklund, A., & Dunnett, S. B. (2007, May). Fifty years of dopamine research. *Trends Neurosci.*, *30*(5), 185–7.

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

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

Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006, Oct). The misbehavior of value and the discipline of the will. *Neural Networks*, *19*(8), 1153–60.

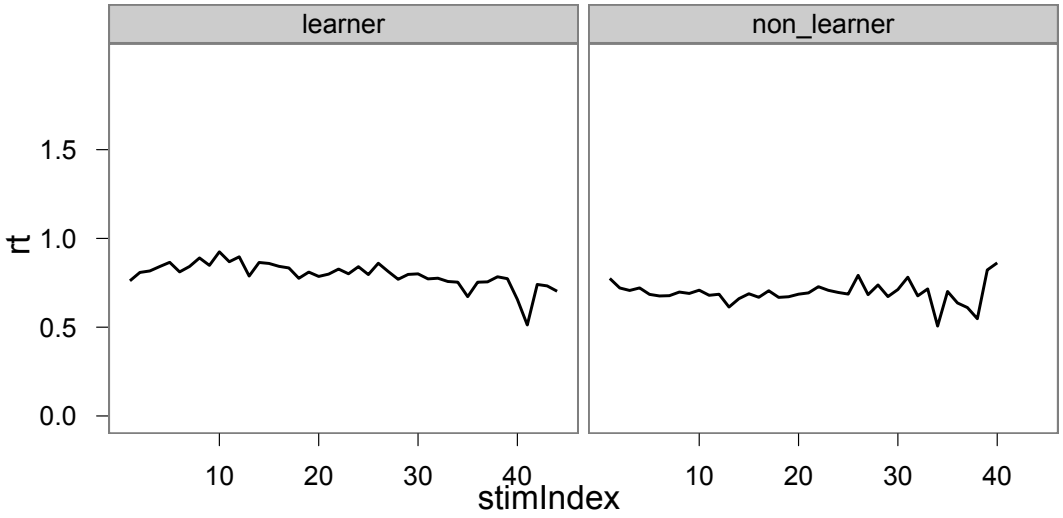

Figure 4. Mean reaction time (black), averaged for all 6 stimuli by trial, blue line and grey represent a linear regression fit of the data and bootstrapped 95% confidence intervals, respectively. See Fig for learning criterion and other relevant details.

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

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.

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.

Menon, M., Jensen, J., Vitcu, I., Graff-Guerrero, A., Crawley, A., Smith, M. A., et al. (2007, Oct). Temporal difference modeling of the blood-oxygen level dependent response during aversive conditioning in humans: effects of dopaminergic modulation. *Biol Psychiatry*, *62*(7), 765–72.

Pessoa, L., & Engelmann, J. B. (2010, Jan). Embedding reward signals into perception

and cognition. *Front Neurosci*, *4*.

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.

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.

Schonberg, T., O'Doherty, J., 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*.

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.