Rewards are categories.

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

Discussion

To review, I wanted to know whether or not cognitive rewards are represented as categories in the human brain. And whether such a representation might impact the reinforcement learning process. To start to answer these two interrelated questions, I had participants complete a stimulus-response task using with pre-trained perceptual categories as rewards, one category for gains and one for losses. And based on the behavioral and neural findings of this work, which I'll now discuss in detail, I conclude that cognitive rewards are indeed represented as categories. And I'll further argue that category representations would be a reasonable mechanistic explanation for the generalization of secondary rewards.

Taking us to can

In the behavioral task reward were literally categories, information integration (II) categories to be specific. II is classic category structure, much studied in humans and other animals (Smith et al., 2011; Ashby & Maddox, 2011; 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, while recruiting procedural memory which relies heavily on the dorsal striatum (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). This lack of verbalizbility and multi-dimensionality make the reward categories irreconcilably different than the classical rewards almost universally used human subjects studies (e.g "Win $1", “Correct!”, “Yes!”). Despite this, participants easily and rapidly learned using the II categories. Performance, as measured by both accuracy and reaction times, were nearly identical to similar tasks using verbal rewards (p**??**).

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

The behavioral and neurological consistency observed in stimulus-response learning using classical and II reward categories means that perceptual categories can act as rewards and so, reversing that logic, rewards *can be* categories. Which leads naturally to the next analysis, whether the same neural algorithm(s) that mediate classical reward learning facilitate the reward categories actions as well.

Reflected in error(s)

Known Pair's Logic. However before making any claims on the modeling data, I need to get some logical preliminaries out of the way. Many of the models of interest are both covariate and dependent. Under generic statistical circumstance it is difficult, or even impossible, to compare the fit of such models. However in limited cases strong, even casual, 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_guass" differ only by the similarity adjustment of the reward (i.e. Eq **??** and **??**). Because both models are fit to the same data[1] and so have identical signal-to-noise ratios, the 1.5[2] fold increase in information that comes from using "rpe_acc_guass" in the dorsal caudate *must* be caused by that single change (?, ?). So while 1.5 would be small increase when comparing two noisy random variates (Anderson, Burnham, & Thompson, 2000; Forster, 2000), I argue that, (1) because uncertainty is constant between the fits, and (2) because we also know the exact relation between two models, and (3) that the models prediction's only sometimes diverge (compare columns in Figure **??**), 1.5 should instead be considered strong evidence when paired-models are compared.

Categories, in all the right spots. Of the regions examined, only the striatum (i.e. the dorsal caudate, putamen, and ventral striatum), along with the ACC and PCC,

Head, ACC, PCC, Insula, mention in passing the PFC.

Then deal with the putamen

[1] Using the same deterministic loss function

[2] Bilateral average

What to say about the rest of cortex?

A fit inconsistency. But what about the "rpe_acc" bieng nearly the best. I argues this is an artifact the dynamic range, solvable either by WM or adjustment.

Making a general sense

Simply stimuli generalize well (by inference, no training is needed) in humans and animals. Mechanistically how this occurs has not been studies. I argue that the even Simple stimuli like tones have categorical representations and that these representations, reflected in the dopamenergic prediction errors, facilate stimulus generalization. Note that the simplest animal generalize on the first new (related) example. This means that categorical representation must be in place prior to that first event. That is, categories are a native representation of the stimuli, not one learned after the fact. And indeed a categorical basis for even simple stimuli is advantageous; due to intrinsic noise in neuronal encoding the same stimulus viewed twice must have a different representation. A categorical representation could (or should in any case) nativly overcome such noise.

Simple rewards decay by exp/or guass too.

Based on .... TODO rewards are categories, or to be a bit less bold, rewards are categories?

Future consequences

Open question is whether rewards are *only* categories. That black cat that crossed your path last Friday, or CATS (the musical), contrasted with category of all cats. Generalizable representations, i.e. categories, are a basic feature. This basic feature extended to rewards. This is not to say specifics do not matter - O'Reilly's trade-off paper. The reward processing and category learning systems share a marked degree of overlap.

The same argument or low outcome repeatability applies to robots or other computational agent just as well. In computer science treating reward category might allow the more flexible and rapid learning, as was the case for my human participants.

Behavioral predictions (however see scaling, WM).

References

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.

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

Cincotta, C. M., & Seger, C. A. (2007, Feb). Dissociation between striatal regions while learning to categorize via feedback and via observation. *Journal of cognitive neuroscience*, *19*(2), 249–65.

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

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.

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., 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.

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.