Chapter 4: Model-based Choice in the Brain

Colin F. Camerer
camerer@caltech.edu
4/1/25 8:47 PM

In the last chapter we learned about some of the multiple neural systems which create choices. One group of such systems, reflexive and prepared, are "factory-installed" and-- by definition—are both mentally costless to implement and difficult to inhibit. The second system is Pavlovian, learning passively (without any choice) to and associate environmental sensory states with rewards. The third system, which is more interesting neuroscientifically, is called model-free. It uses principles of active reinforcement learning to associate actions with rewards. We learned that a lot is known about the neural circuitry of model-free learning through reward prediction error (RPE), including studies that measure RPEs rather carefully with methods ranging from fMRI BOLD signals to voltametric measures of dopamine.

This chapter is focused entirely on a fourth system. This system is called model-based or goal-directed valuation. (We'll use those terms synonymously.) There are at least two simple ways to differentiate model-free and model-directed systems intuitively. One way is that highly automatic "overlearned" habits are model-free, not model-based. They work as if actions themselves have acquired value (decision utility) even if the actions do not accomplish the goals

they normally do. In model-free choice there is no model that dictates how model-free action reward predictions should be altered when environmental cues change.

An American travelling to England for the first time, trying to cross a street, looks to the left for cars coming in the right-hand lane. The "look left" action is automatically valued. It is model-free. But in England cars drive in the left lane, so Americans need to look to the right for the most dangerous oncoming cars. Looking left is model-free; it is not goal-directed or model-based, because it does no adjust to achieve the goal of avoiding an accident.

Model-Based Goal-Directed Valuation

The term "model" used in neuroeconomics is a cognitive representation of how different actions lead to valued outcomes. Such models can range in simplicity to complexity. Here are some examples:

- estimating calorie-taste values of all the sandwich items on In-and-Out Burger menu (there are only three in the regular menu);
- the mental version of a Google Map showing the different routes from Caltech to Manhattan Beach (with red marking route segments with heavy traffic);
- an elaborate spreadsheet estimating the value of a startup, which is daydreaming about becoming an initial public offering (IPO).

Historical Background

The modern distinction between habits vs goal-directed behavior, and later model-free vs model-based valuation, is the resolution of a lively debate from the first half of the 1900's about what animals are learning when they seek and get reward.

The earliest view in the debate was exemplified by "the law of effect" (Thorndike, 1911)—namely, animals learned to repeat a rewarded response after a stimulus, a process of instrumental active learning sometimes known more specifically as S-R-O (stimulus-response-outcome) learning. In this view, animals build up a workplace manual of what R should follow each S, with minimal aggregation of similar stimuli into generalizations.

"Behaviorism" became the name for the school of thought held by strongest believers in the model-free S-R-O view, from Edward Thorndike to John Watson. They felt that all behavior— even in humans— might be explained by reinforcement learning processes. For example, Bertrand Russell (1921, p 37) wrote:

> Or you make a model of the Hampton Court maze, and put a rat in the middle, assaulted by the smell of food on the outside. The rat starts running down the passages, and is constantly stopped by blind alleys, but at last, by persistent attempts, it gets out. You repeat this experiment day after day; you measure the time taken by the rat in reaching the food; you find that the time rapidly diminishes, and that after a while the rat ceases to make any wrong turnings.

So far Russell hasn't said anything controversial. Many maze learning experiments do just what was described. The rats asymptote at near-perfect performance of nothing in the maze changes over days of learning. Then he goes further:

> It is by essentially similar processes that we learn speaking, writing, mathematics, or the government of an empire.

Well now! That is a remarkable claim indeed. There some truth in Russell's grand behaviorist claim, but he goes too far (depending, of course, on what he means by "essentially similar").

While Thorndike and later behaviorists were trying to show how much learning could be explained by simple conditioning, Tolman (1948) proposed a "purposive behaviorism". His idea was to integrate the undeniable principles of S-R reinforcement learning with the idea that animals also created a "cognitive map" which could be used to achieve new goals that were *not* previously learned by direct S-R. 4.1

The opposing Thorndike and Tolman views now survive in the form of an accepted co-existence of model-free (Thorndikian) and model-based (Tolmanian) systems.

A beautiful experiment establishing the kind of cognitive maps Tolman believed in is illustrated in (Figure 4.1). In the experiment rats learned to travel through a maze for food (Figure 4.1a). Then they were placed in a more complicated maze and their previous travel path was blocked (Figure 4.1b). Even on the initial trial, the rats used their knowledge of the physical location of the food from the previous maze—independent of the learned sequence of actions to reach the food, which were not available in the new maze-- to choose a brand new path. The most common path the rats chose, on the first trial in the new maze, put them very close to the geographical location of where food was delivered previously. It's as if they have an internal GPS reference system independent of the maze structure, and follow it when a new maze is presented.

Tolman and others who believed in his view (called "Tolmaniacs") discovered several other robust facts about animal behavior which were not easily reconciled in the simple S-R view.

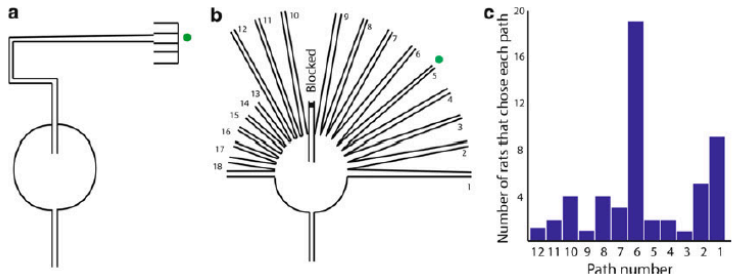


Figure 4.1 An early experiment used by Tolman (1948) to show that rats can build and use spatial representations. (a) Rats were first trained to find a food source located below the *green point* (a light). (b) After training, the rats were placed in the same starting position at the bottom of the maze, but found their usual route blocked. Instead, they now had multiple alternative arms

they could run down. If the rats had learned only an association between the original stimulus route (in (a)), action, and reward, they would never reach the food at the end of the new path (6). (c) Histogram of the frequency with which rats first chose to run down each of the many arms. They did not choose paths 9 or 10 often, although those paths are physically close to the blocked path. Instead, they most often chose path 6. Source: Huys et al (2014).

Painstaking research showed several effects consistent with the buildup of internal cognitive maps in maze learning. These include: Latent learning, VTEs (vicarious trial-and-error), and sensory preconditioning. Each of these will be explained next.

Latent learning

Rats that are allowed to freely explore a maze, with no reward, learn faster when reward is introduced (compared to rats with no free-exploration experience). The rats seem to be creating memories about the maze pre-reward and using their simple map to learn faster when cheese is on the line (Blodget, 1929; Thistlewaite, 1951). This is called "latent learning" because it is not learning created by rewarding experience.

An example of latent learning is called the "irrelevant incentive effect" (Kreickhaus and Wolfe, 1968). A typical example uses rats that are thirsty but not salt-deprived. The W(ater) group learns to lever-press for water, and the S(odium) group learns to lever-press for sodium solution. Both groups work equally hard (i.e., press levers at similar rates) since plain water and sodium solution are about equally rewarding.

Later, all the rats are salt-deprived and they have a fresh chance to lever-press for liquid. The S group, who had previously learned that their liquid reward contained salt, pressed more vigorously than the W group. *They had apparently noticed sodium in the water, even when it wasn't especially valuable for them during the first training phase*. They remembered this fact, and it influenced their vigor in trying to reach a new goal— the goal of getting more salt into their system because they were salt-deprived.

In economic terms, when the marginal utility of salt went up in the S-group rates (because they were salt-deprived by the experimenter), they drew upon their knowledge that the liquid they previously drank had extra sodium and responded more vigorously. In psychological terms, This behavior is called "vicarious trial and error" (VTE; Muenzinger, 1938). Rats who perform more VTEs also learn to reach new reward goals faster. The rats are building up a cognitive map by VTE-ing around. VTEs are a primitive relative of "fictive learning", which plays a central role in human learning (see Camerer and Ho (1999), Zhu, Mathewson, and Hsu (2012) and Appendix A).

although the sodium content in the S group was *not* rewarding during the training phase, knowledge of it was used to compute an increased goal value for the sodium solution, leading to an enhanced goal-directed response (compared to the W group).

VTEs

While rats are learning to navigate a maze for a food reward, they often pause and look around. (In military exercises, outdoors activity in remote locations, and other kinds of activities people are also trained to look for physical cues and landmarks, to be able to retrace their steps.)

Neural evidence for the existence and nature of VTE's comes from activity in the hippocampus (a region that consolidates memory and also generates prospection about the future). Hippocampus contains very localized "place cells"—neurons that only fire, after substantial learning, when the rat is at particular places in a maze. Furthermore, there is evidence that during VTE exploration, there is firing of place cells corresponding to locations the rat is likely to go in the next second or two, as if the rats are "planning ahead" about what could happen if they move in different directions, and the planning is expressed by firing of soon-to-be-reached place cells (van de Meer, Redish 2009). (To an economist, hippocampal place cells firing during maze navigation are an intriguing implementation of a mental representation of the continuation value of moving ahead various routes, as is used in dynamic programming .[93])

An important piece of additional *causal* neuroscientific evidence is that lesioning hippocampus reduces VTE behavior (Hu, Amsel 1995). This shows that VTEs are encoded in hippocampus, which is a biologically plausible location (because hippocampus is involved in memory storage, consolidation, and retrieval) if the VTEs are used to build up a memorable cognitive map of the maze.

Sensory preconditioning

Another example of learning beyond simple reinforcement is "sensory preconditioning". To establish this effect, an animal learns an association between two cues A and B: When A is presented it is always followed by B. This link is learned through stimulus-stimulus association. Then, a new relation between cue B (presented by itself) and a reward is learned, in a passive model-free way: That is, the presence of B always predicts reward, and the animal learns that association.

Now what happens when the sensory cue A is present? Cue A has *never* been associated with reward, so there is no reason for a reinforcement-learning animal to respond to A in a Pavlovian way, as if predicts reward. However, if animals naturally couple the learned association between A and B, and the passively conditioned association between B and reward, then animals *will* expect reward when A is presented.

Indeed, animals *do* seek reward in response to the cue A, an empirical fact that is called sensory preconditioning. (This interesting phenomenon will emerge again in chapter 4 as a driver of learning to value 'new products' linked to other valued products through S-S association.)

The Tolmaniacs' view was that this tiny chain of logic requires a minimal kind of cognitive map, in which A, B, and reward are linked so that the A-reward association is inferred from the combination A-B association and B-reward.

BOX: Exploration and exploitation in repeated choice

---

[93] Dynamic programming is a method of solving for an optimal policy of sequential choices, where each choice at a decision node in a sequence creates a reward and also transitions the available choices from one set to another. If the sequence ends at terminal state S_T, then the value at the last decision node is used to compute the possible values from each choice made at node T-1, taking into account the value of whatever decision node follows the choice made at T-1.

A subtler element of goal-directed choice is the "exploration-exploitation tradeoff". The term comes from analysis of what are called "multi-armed bandit" problems (a "bandit" is an old-fashioned term for a rewarding slot machine). In a bandit problem, there are multiple choices, their underlying reward distributions are not fully known, and an agent makes one choice each period (Gittins (1979); Robbins (1952); Thompson (1933)). Choices are made repeatedly from the same set of bandits over some expected horizon of repeated choices.

People who choose bandit A earn an immediate reward **and** also learn about the distribution of the bandit A rewards. The choice creates reward plus information value. Optimal learning should therefore be guided by a combination of expected reward and expected information value. The latter is often called an "exploration bonus".

In a new bandit situation, the exploration bonus is high. The exploration bonus is usually increasing in variability. The key to good bandit choice is to adjust the proper balance of exploiting historical reward information and acquiring new information. Optimal agents should explore more persistently if there is more uncertainty and a longer time horizon of choice.

Bandit problems are more difficult to solve exactly than maximization with known reward distributions.[94] However, it is conceivable—and I think, likely— that humans are rather good at making the proper exploit-explore tradeoff. Why? In "restless" bandit problems the underlying reward distributions are nonstationary. Because of the nonstationarity, in optimal algorithms there should be some degree of periodic exploration and change-detection. For repeated choices people make, such as commuting, workplace decisions, food, entertainment, etc., this is arguably the canonical type of decision people face with mean-reverting rewards. The limited literature on lab experiments suggests that human decisions are = close to optimal algorithmic performance (e.g., Banks, Olson, and Porter (1997); Lee et al. (2011)).

END BOX

Model-based Choice

The previous subsections laid the foundation for understanding model-based choice. Even in the starkest behavioral conditions, animals such as rodents, have abstract mental representations that enable them to compute action valuations, even for actions they have not executed previously. Henceforth we will call these representations "models" (or "world models" in computer science language).

---

[94] Whittle (1979) wrote that "[the bandit problem] was formulated during the [second world] war, and efforts to solve it so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage." (TO DO can't find direct source)

The distinction between model-free and model-based value computation is fundamental. ("Goal-directed" overlaps with model-based for our purposes in this chapter, because the model is calculating a goal value.) It is useful theoretically, is well-supported in neural data, and is promising as a way to characterize differences in individuals as well as comparative static responses to external constraints. It is the biggest potential gift, so far, from neuroscience to rational choice social science. Figure 4.2 illustrates the difference between model-free and model-based.

Model-free evaluation is mentally easier than model-based valuation. All that's required in model-free evaluation is to store (or "cache", in computer science jargon) a numerical action value. When that value is updated according to new experience, the updated new value replaces the old value. This updating process can usually be expressed in a simple learning equation. This type of model-free learning appears to correspond to behavioral results from active learning, which has been implemented experimentally with very reliable replication over a range of decades, in an enormous range of species.

In contrast, model-based evaluation requires a mental "model" representation, prescribing what may follow from choosing an action under particular conditions (state variables). Note that the forecasted values embedded in the mental model could be based on memory, could be instructed formally, could be socially learned by observing others, or could come from an explicit abstract model as in constructing a decision tree. Memory, instruction, and social learning will be discussed in the next chapters 5-6.

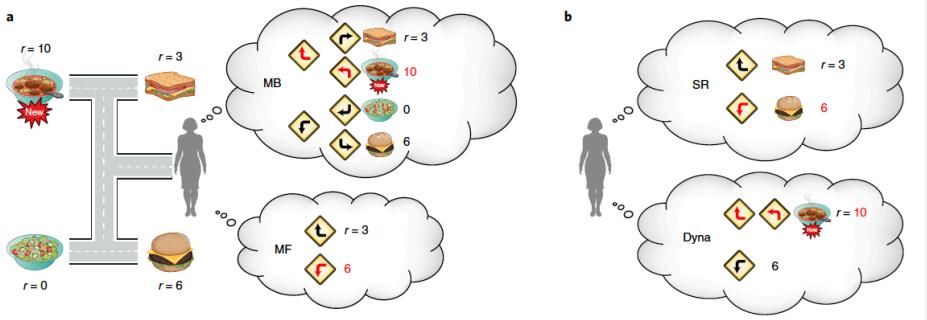


Figure 4.2: Model-based (MB) and model-free (MF) evaluation illustrated by choosing different routes and foods for lunch. (a) In the bottom thought bubble, the decision maker has learned model-free values of the actions of turning right (r=3 from a sandwich or turning left (r=6). In the model-based top thought bubble, she also knows there is a new, but not-yet-experienced soup shop with a high reward r=10. (b) Graphic illustrates how two variants, successor representation (SR) and Dyna encode actions, foods, and rewards a little differently. Source: Daw (2018) Figure 1.

A simple way to distinguish between model-free and model-based valuation is to create an environment in which a choice can be rewarding while *also* providing information that

another choice would be even better. If learning is purely model-free, the rewarded choice will be repeated. If learning is model-based, the rewarded choice will not be repeated; the even-better other choice will be made instead.

A task precisely designed to distinguish model-free and model-based choice was introduced by Nathaniel Daw et al (2011) and has been used many times with interesting results. It's called the "two-stage task". The structure is shown in Figure 4.3.

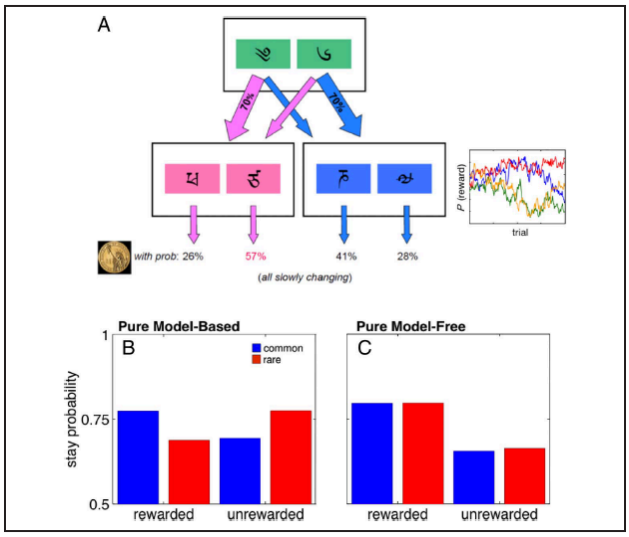


Figure 4.3: (A) In each trial people one of the two green symbols (first stage). Each choice leads to a second-stage choice between one of two pink (left) or blue (right) symbols. Crucially, the left green symbol choice is more likely (70%) to lead to the left pink choice pair and less likely (30%) to lead to the right blue choice pair. The opposite probabilities hold for the right green symbol choice. These two transitions from each of the first-stage choices are called "common" and "rare" transitions. Each of the second-stage choices has a probability of a fixed reward—in the example the probabilities are 26%, 57%, 41% and 28%. These probabilities slowly "drift" and change value (according to the inset box time series). (B) Model-based learning uses learned information about the transition probabilites and rewards. The highest likelihood of "staying" with the last choice is when the transition was common and rewarded, or rare and rewarded. That is, when the rare transition was rewarded by a second-stage choice, a model-based thinker realizes that she should switch to the other first-stage choice which makes the opportunity to make the second-stage choice more often. (C) Model-free learning prescribes a higher probability of staying with the last choice when the outcome was rewarding, regardless of the transition. Source: Doll et al (2014).

At the first stage, subjects choose between one of two stimuli, which lead "commonly" (most often, with 70% probability) to two green second-stage choices. The first stage choice could also lead, with a lower "rare" probability (30%) to two blue second-stage choices. The

first-stage choice can be thought of as a “mostly-green” or “mostly blue” choice. Each of the second stage choices gives either nothing, or a fixed reward, with a hidden choice-specific reward probability that is constantly changing over time (as shown in the figure inset).

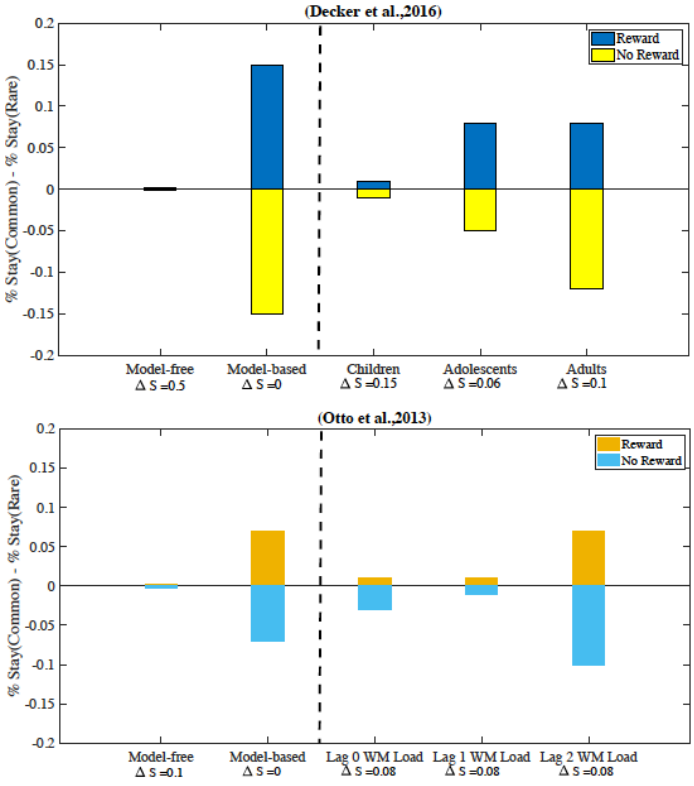
Recall that the four second-stage choice-specific reward probabilities are drifting over time. Therefore, at any given time, either the green pair or the blue pair is a “better” pair of choices to be able to choose from, because the best of one pair has the highest current reward probability. Suppose the subject picks the mostly-green first stage choice. Then suppose the rarer event occurs, so that she ends up with a choice between two blue choices, and she wins and gets a reward.

If learning is model-free, she learned an association between choosing mostly-green and reward, so she will choose it again—we will say her P(stay) (“staying with” the mostly-green left choice) is high. If learning is model-based, however, she will realize that the reward she actually received nudges up the subjective belief that blue choices are good choices to have. And she *can have* blue choices, more of the time, if she switches to the mostly-blue choice at the first stage.

So a model-based learner has a lower value of P(stay) in the rare, reward condition. A parallel argument applies if the choice is not rewarded: Model-based learning implies switching if the choice is not rewarded and the transition was common. The resulting profiles of P(stay), depending on whether the first stage choice led to the common or rare branch, and whether there was a reward or not, are shown in Figure 4.3 (B,C). Readers, if you haven’t seen this before, sit with it for a minute. It is a good company way to appreciate the model-free vs. model-based distinction.

The two-stage task has been used in many studies. To illustrate many findings compactly and visually, we will use a new method shown in Figure 4.4. The y-axis is an estimate of P(stay|common)-P(stay|rare). In the left two columns (marked by a dotted vertical line on their right) shows predictions about this statistic from model-free and model-based learning. The model-free prediction is always that P(stay|common)-P(stay|rare) is zero because staying only depends on reward for the first-stage choice, not on whether the common or rare transition occurred. (Intuitively, the first-stage gets the ‘credit’ for reward even if it came from the rare second-stage choices.) The model-based prediction *does* depend on whether the trial outcome was rewarding or not. If it was rewarding P(stay|common)>P(stay|rare) because reward makes P(stay|rare) low—after a reward at the second stage, a model-based person switches to the opposite first-stage choice which makes that type of reward more likely. Similarly, if there is no reward then P(stay|common)<P(stay|rare) (the difference on the y-axis is negative) because staying after the common second-stage consequence is a bad idea.

Visually, the point is simple: If behavior is more model-based there will be larger positive and negative bars in response to reward and no-reward in the actual empirical frequencies of staying. Tell your eyes: Big bars= model-based.


(Decker et al.,2016)
Reward
No Reward
% Stay(Common) - % Stay(Rare)
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
Model-free
Δ S =0.5
Model-based
Δ S =0
Children
Δ S =0.15
Adolescents
Δ S =0.06
Adults
Δ S =0.1
(Otto et al.,2013)
Reward
No Reward
% Stay(Common) - % Stay(Rare)
Model-free
Δ S =0.1
Model-based
Δ S =0
Lag 0 WM Load
Δ S =0.08
Lag 1 WM Load
Δ S =0.08
Lag 2 WM Load
Δ S =0.08

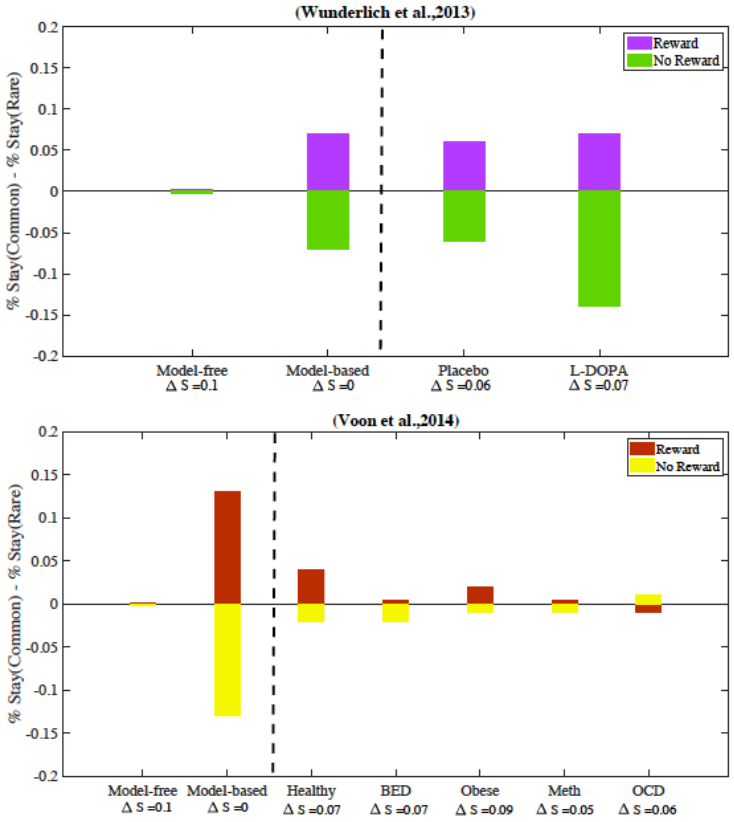


Figure 4.4: Statistics showing conformity with model-free or model-based behavior. Each column and bar pair illustrates three numbers for an experimental theory or treatment: (i) The %Stay(Common)-%Stay(Rare) after reward (positive values); (ii) %Stay(Common)-%Stay(Rare) after no reward (negative values); (iii) the difference DS= %Stay(Common)-%Stay(Rare), plotted below the x-axis labels. The left two columns indicate stylized predictions for model-free MF and model-based MB. For MF the bars plotting %Stay(Common)-%Stay(Rare) are zero and DS is substantially positive. For MB the bars plotting %Stay(Common)-%Stay(Rare) are positive after reward and negative after no reward, and DS=0. The four panels show: (A) Progression from MF to MB with age (Decker et al 2016); (B) immediate working memory load in lag-0 and lag-1 trials makes behavior close to MF (Otto et al 2013); (C) Parkinson's patients are more MB under L-Dopa treatment (Wunderlich et al 2013); (D) Neuro-healthy controls are slightly MB but other disorders are close to entirely MF (Voon et al 2014). Source: Original graphic, Xiaomin Li.

Figure 4.4 shows statistics based on frequencies of P(stay|rare,r=1) and P(stay|rare,r=0)—i.e., good stays and bad stays—in several different studies. The initial behavior, reported in Daw

et al. (2011) (not shown) shows behavior modestly in the direction of model-based learning. Children (ages 8-12) (Decker et al 2016) and people under cognitive load (Lag 0 and Lag 1 WM Load, Otto et al 2013) are more model-free. Administration of L-DOPA to Parkinson's patients improves model-based learning (Wunderlich, Smittenaar and Dolan, 2012). Several impulse-control disorders are associated with more model-free learning (Voon et al 2014).

This simple distinction between good and bad staying is of huge empirical importance. A mountain of studies going back to the 1960s have documented evidence for a "win-stay, lose-shift" (WSLS) heuristic. In the two-stage task, WSLS is closely related to model-free learning. WSLS ignores information from a win about whether switching could be better, and ignores information from a non-reward "loss" about whether switching might be better. Unfortunately, since the Daw et al. (2011) paradigm is rather new, most of the historical studies showing WSLS do not even consider the value of the opposite pattern or its modulation by structure.

Recently Akam et al (2015) and Kool, Cushman and Gershman (2016) pointed out a substantial flaw in the two-stage paradigm: Increasing the weight *w* on model-based valuation does not actually create higher earnings. And when earnings are regressed against the simulated behavioral model parameters (the learning rate and the inverse temperature), there is a "flat maximum" in which many parameter combinations lead to very similar earnings. As a result, this is not an ideal task to measure the tradeoff between performance and cognitive difficulty, which is thought to be a key feature of when control is adjusted between model-free and model-based valuation. Kool et al. (2016) proposed and tested a different two-step task which is a little simpler to comprehend and more discriminating It uses a higher reward drift rate and continuous non-binary rewards. In an MTurk sample, they get a weight on model-based computation of

Rethinking model-free and model-based: Scientific dichotomies usually creak and shift under the weight of evidence and then give rise to better classification (shifts of geological tectonic reshaping countries and continents—sloooowly-- seems like an apt metaphor.) Many scientists have suggested ways to improve upon the conventional use of MF and MB. Daw (2018, pl. 1499) wrote that "In the end, perhaps we are not creatures of two minds—or three, or four". Daw particularly noted the role of memory, which we'll discuss in the next Chapter.

Feher da Silva et al (2020, 2023) compiled the most provocative and compelling evidence that the MF-MB dichotomy, as conventionally used, is missing important ingredients and has not passed appropriately stringent tests.

The authors' simplest empirical contribution is that when the two-stage decision problem is described with a narrative "cover story" explaining the common and rare transitions, behavior becomes sharply more model-based. In one version the initial two states are "magic carpets" that are more likely to go to one mountaintop than another. At each mountaintop they choose one of two lamps, from which a genie (creating reward) comes out. The narrative explanation is that erratic winds sometimes push the carpets off course to the rare mountain. Another example is spaceships which sometimes go to the rare unintended planetary targets

The authors suggest that while accurate MB behavior is easy to characterize, there may be a lot of other non-accurate-MB behavior which is not truly MF (because it uses ad hoc and varying models that are not accurate). In an MF-MB classification all these residual behaviors are likely to get dumped into a MF bin.

In the two-stage task, after the first choice is made the transition to the second state is revealed, which in MB analysis should create a "state prediction error" (and an associated

change in reward prediction). There is a second prediction error after the outcome of the second state choice is revealed. Remarkably, in a large sample of about 100 subjects, there are no RPEs evident in basal ganglia during the second stage interim state, although there is a strong response to reward feedback (but not RPE).

The authors' results suggest a foggy, but important, pivot for scientific progress: Treat pure MF choice as a tentative process that is not yet nailed down by the workhorse two-stage task in either behavior, or as effort-reducing (which they measure using pupillometry) or in fMRI. There may be more exploration among variants of MB learning (often misestimating the actual model structure) that is being swept up, mistakenly, into the MF category.

In an April seminar at Caltech, I suggested to Todd Hare that one way to learn about what was happening in the possibly-catch-all MF category, because there is so much data available from the two-step task, was to use machine learning to extract cognitive strategies. Ji-An, Benna and Mattar (2023) had done just that! They found that "tiny" recurrent neural networks ("tiny meaning with a small number of layers in an RNN) were able to detect cognitive strategies that are certainly not model-free in a variety of data, fitting better out-of-sample than the simple MF-MB decomposition. There is only a small caveat, which is that the RNNs have more free parameters so only work well with 500-2000 trials of training data. But such data sets are well within our reach, and the authors' idea will certainly be fruitful in moving beyond the MF-MB dichotomy by revealing a variety of other strategies.

Planning It's time to introduce the concept of a decision "policy". A policy is a complete description of what to do at all possible choice sets (also called "decision nodes"). (Model-free is a choice policy that associates a cached reward prediction or value with each state and action and chooses actions accordingly.) Intuitively, a business manager could write down a policy, go on vacation, and leave an assistant to execute the policy. Later we will discuss how policy evaluation and learning might occur, which transcends the simpler types of model-based learning.

Model-based processes are often described as planning, which anticipates (future actions and outcomes (such as a decision tree). How do we know that the model-based process actually involves *planning ahead*, in the sense of prospectively considering future events or choices? A beautiful study by Doll et al (2015) provides a rather clear answer, and a method that could be used in many other studies involving mental planning.

Their design involves two sequential choices. The first choice is between either two faces or two tools; whether the Face or Tool sets pops up is randomly determined, out of the control of the subject. For each of those Face or Tool pairs, picking one of the faces or one of the tools always leads to a second choice set, consisting of two body parts or two scenes. It is those second-stage choices that lead to probabilistic rewards. Reward probabilities for each of the four possible second-stage choices drifted slowly between .35 and .75 over 272 trials.

The key feature of the design is that value should continually update based *only* on which of the four body parts or scenes are chosen in the second stage. The first stage choice is inconsequential, because either of the two Body and Scene choice sets can be deterministically

reached by picking the corresponding face or tool. In other words, if the tracked value of the first scene is the highest, you can always reach it by choosing the face or tool that triggers the Scene set.

However, model-free learning crudely associates all stimuli with value, including the first-stage choices. If the last trial started with the Face set, and the choice led to a high-value Scene choice that paid a reward, the model-free learning will increase the perceived value V_{t+1}(Face).

Doll et al. first fit model-based and model-free learning models to the choices each person made. As has been seen many times, the overall fit is roughly halfway between model-based and model-free, and subjects vary systematically in the extent to which they appear model-based.

The next trick exploits the fact that visual images of body parts and scenes activate different brain areas. They used a "functional localizer" method, which consists of simply presenting only the images over many trials (not linked to any choice), and seeing what brain areas are activated by the different image sets (e.g. Body areas vs. Scene areas) separately for each person. With these filters in hand, they then looked for activity in the localized areas during the *first* choice stage.

Suppose subjects are planning ahead, planning to choose the high-value scene, as they ponder which Face choice they have to make for the Scene set to appear. Then during the Face choice, the brain areas localized for the scenes should be active in the "mind's eye". The key assumption here is that actually seeing a particular scene, or simply planning to choose it when it shows up in a few seconds, will activate overlapping areas.

That type of revealed planning is just what they found: The subjects who acted more model-based, as revealed by choices, had stronger activation in the localized brain areas associated with later second-stage choices, during the very first stage. Planning involves "internal seeing".

To summarize, current research clearly establishes neural correlates and behavioral expression of two systems, respectively model-free and model-based. Encoding prediction errors used to implement TD-learning in model-free valuation, and hierarchical prediction errors (e.g. for states) have been established in regions including VStr bilateral intraparietal sulcus and lateral PFC. But how exactly does the model-based system construct an estimate of value by combining these various inputs?

To answer this question is it useful take an elevator up one floor to a computational level that abstracts from neural detail. The discussion of model-free and model-based systems earlier focussed on the special case in which there is often a set of actions and a limited number of states. The Daw two-step task, for example has two starting states (the initial choice set) and two final states (the second-step choice pairs). In two-step tasks, and much more generally, one can define state-action sequences (usually called "policies" π), and try to understand learning about policy values rather than action values. To nail down the concept of a policy, note that in the

two-step task there are eight possible policies—for each of the two initial choices, there are four possible second-step choices.

The idea is that the brain uses gradually-accumulated knowledge of structure to generate predictions of how rewards result from policies. Based on knowledge about rewards, a learning process leads to an estimate of the likelihood of reward r from a policy π, denoted p(r|π). "Inverse inference" or "generative modelling" is then performed, using Bayes' rule, to decide how much weight to put on policy π to earn the highest-valued reward, p(π|r*). The process can then be iterated, either learning or simulating paths by imagining what policies will lead to, in order to build up a more precise estimate of the policy value p(r|π). In simple cases it can be shown that such an algorithm converges to choose the optimal policy.

An early use of generative came from the brilliant insight of Helmholtz (1860/1962) about how vision works. Helmholtz suggested that vision is not a snapshot that is recording exact details. Instead, vision is a rapid unconscious inference of what is out there. Vision is a *guess* that sheds nuisance information to extract gist

To express Helmholtz's idea in modern language, knowledge in the brain encodes a conditional probability distribution, p(visual image|actual scene)—that is the generative model. The generative model is an internal guess about how well actual scenes are mapped into internal images. The guess about what a person saw— the posterior probability p(actual scene|visual image)-- is made by Bayesian-inverting the generative model p(visual image|actual scene) using Bayes' rule. This general model is now well-supported (Dayan et al., 1995; Friston, 2005; Knill & Pouget, 2004; Mumford, 1992, 1994).

Solway and Botvinick (2012) implement a generative-model structure as a theory of goal-directed learning. The intuitive idea is that people construct evidence about the rewards that result from action sequences (policies) p(r|π) by iterated exploration. Then this generative model is Bayesian-inverted to adjust likelihoods of policies, for choice, p(p|r). Their approach generates some nonobvious predictions:

1. There should be immediate coding of both early and final goal states, rather than ramping up to final goal states throughout a task (Saito 2005)
2. There should be representations of expected reward at each stage, working through a tree. These are known as continuation values in dynamic programming.
3. In the simplest view, action values V(a) are thought to be represented in a habit system. However, in the generative view V(a) are inferred backward from state value and an action-state transition function. As a result, a disruption of the state value evaluation will also disrupt computation of V(a).

Interestingly, the hypothesized generative process inadvertently creates an "optimistic bias". Since the process is designed to find the best policies, it is initialized by weighting outcomes by reward r=1. As a result, inferred generative models of what to do—and what is expected—based on those posteriors will also be too optimistic. Solway and Botvinick (2012, p. 131) suggest that this type of optimism might be a computational underpinning for preferences for control (*a la* illusion of control). The idea is that when people are freely choosing, the optimistic evaluation will emerge as a byproduct of how learning works, but if the same choices are made by an outside agent there is no such homegrown optimism. The theory also implies that

optimism bias will shrink as organisms learn about goal-directed valuation. Optimism should also be present in all species that learn in the way they hypothesize (which draws heavily on work with rats in mazes, and primates). As a result, it could be that human optimism does not require special abstract human-only concepts of wishful thinking about the future, or a belief-based utility for optimistic beliefs.

Value computation in vmPFC

Earlier we discussed how dopamine neurons in midbrain, striatum, and their projections elsewhere are important for encoding reward prediction error used in TD learning. Besides those regions, another central region commonly involved in valuation and choice is ventromedial prefrontal cortex (vmPFC).

First note that vmPFC is not a distinct anatomical region; indeed, the term is used in different ways to describe a collection of regions and different scientists often use different boundaries (Yu, 2018). One popular classification of brain regions was a division into 52 areas proposed by Brodmann in 1909, based on staining of cells and differentiation of brains region based on cytoarchitecture (cellular composition).

In fMRI reporting, the vmPFC is used to describe the region from anterior cingulate cortex (ACC) below the corpus callosum (the ventral part of Brodmann areas BA 24,25,32), the frontal pole (BA 10) and medial orbitofrontal cortex (mOFC) (BA 14). Lesion researchers usually include central and lateral orbitofrontal cortex (BA 11, 13, 12/47). (We will see quite a few graphs showing regions of vmPFC; it is not very important to remember these BA numbers in absorbing what those graphs show.)

Regions of vmPFC and orbitofrontal cortex (OFC) are thought to have two or more distinct connections or networks—a medial network and an orbital network. Ongur and Price (2000) distinguish the medial network (ACC, vMPFC and mOFC), which is connected to amygdala, entorhinal cortex, and hippocampus, and projects to ventral striatum. A different orbital network including central and lateral OFC, is connected to sensory inputs and is also connected to perirhinal cortex and central striatum. The two medial and orbital networks also connect to each other (through BA 13-14). Ongur and Price (2000, Figure 5) shows the different networks.[95]

vmPFC is in a particularly suitable position to represent valuations because it is connected to emotion and memory centers, as well as reward areas like the ventral striatum. vmPFC is sometimes considered a “common currency” computational region, like a foreign exchange window in an airport where all different currencies are valued accurately and can be exchanged (Levy and Glimcher 2021). More recently, it has been shown that vmPFC indeed represents a domain-general value signal: That is, lots of different object subjective values and represented in overlapping regions of vmPFC. In contrast, central OFC—between mOFC and the most lateral part of lateralOFC-- represents identity category-specific value: That is, different

[95] Clithero and Rangel (2013) report coactivation of other regions with three distinct peaks in mPFC (all part of the putative medial network).

object categories are represented in different regions of central OFC (Howard, Gottfried, Tobler, & Kahnt, 2015; Howard & Kahnt, 2017; Pegors, Kable, Chatterjee, & Epstein, 2015).[96] Newer studies can do even better, within spatiotemporal limits of whole-brain fMRI, in isolating areas of valuation. For example, Suzuki, Cross and O'Doherty (2017) found areas of OFC encoding different aspects of nutritional food values.

Two meta-analyses of valuation experiments provide solid evidence of the general role of vMPFC in value encoding.

Clithero and Rangel (2013) collected peak coordinates from S=81 human fMRI studies in which whole-brain results and parametric contrasts (the *amount* of value) were reported. Clithero and Rangel (2013) focus only on positive encoding of stimulus value (SV).[97]

Figure 4.5 shows regions differentially activated by an increasing *amount* of value either at the time of decision or time of outcome. The main regions, vMPFC, posterior cingulate cortex, and striatum, are seen in study after study. There is also quite a lot of overlap between decision and outcome valuation. This is important because of the discussion previously in this chapter about wanting (decision utility) and liking (experienced outcome utility). If common regions are activated by these two types of utility, that strengthens the hypothesis that wanting and liking are likely to match up. Put differently, if wanting and liking are systematically different, that difference would seem to require specialized processing in some brain regions that encode wanting or liking separately, and not both. For example, imagine a disorder such as Tourette's syndrome which creates uncontrollable "tics", body movements, or verbal exclamations. It is plausible that the disorder makes the brain "want" to exhibit tics, but a person does not subjectively like them.[98]

[96] An intriguing recent theory about vMPFC is that vmPFC navigates through physical or conceptual space using a grid-like coding much as the entorhinal cortex does for spatial navigation—that is, choice objects which are conceptually similar are encoded in adjacent regions of vmPFC (e.g. Hafting ET al 2005) . Does vMPFC actually have such grid-like coding? A study by Yu et al (2018) suggests the answer is No, but many more studies are needed to be more conclusive about this intriguing grid-code hypothesis.

[97] They also did an image based meta-analysis (IBMA) which uses the entire whole-brain SPM maps from particular GLM contrasts, for 21 studies. In the time when these meta-analyzed studies were not done, entire SPM maps were generally not made available, but having those is ideal. The results of the IBMA are similar except for consistent activation of superior frontal gyrus (SFG, -18, 42, 45).

[98] To be clear, I don't know enough about Tourette's or the felt experience of people with the disorder. It is just a dramatic and plausible example of how wanting and liking may be persistently decoupled.

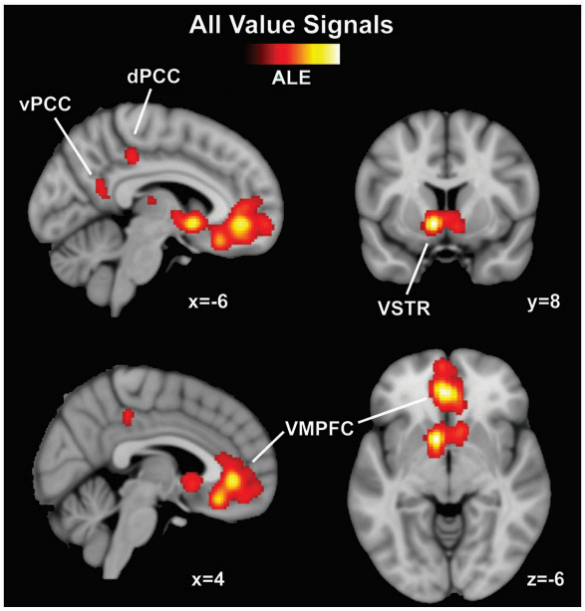


Fig. 4.5. Neural regions encoding positive subjective value. Coordinate-based meta-analysis (CBMA) was used to identify clusters associated with increased subjective value. The CBMA method found distinct clusters in VMPFC, Vstr and both dorsal and ventral PCC (posterior cingulate cortex). The global maximum ALE value was located in left Vstr; this means that the region which was most likely to be activated by increased subjective value, taking all studies together, was lVTstr . The (unlabelled) scale reflects ALE values which range from voxel-level significance of $P < 0.001$ (uncorrected) and a cluster-corrected threshold of $P < 0.05$. The colorbar spans ALE values of $18.x10^{3}$ (min) to $68.89x10^{3}$ (max). Source: Clithero and Rangel (2013)

A meta-analysis published in the same year as Clithero and Rangel by Bartra et al (2013) focused on some of the same questions, as well as some different ones. Their final corpus was S=206 studies and included analyses in which the value contrast could be binary (e.g., high vs. low positive amounts) (Figure 4.6), rather than continuously graded by the amount of subjective value as in Rangel and Clithero.

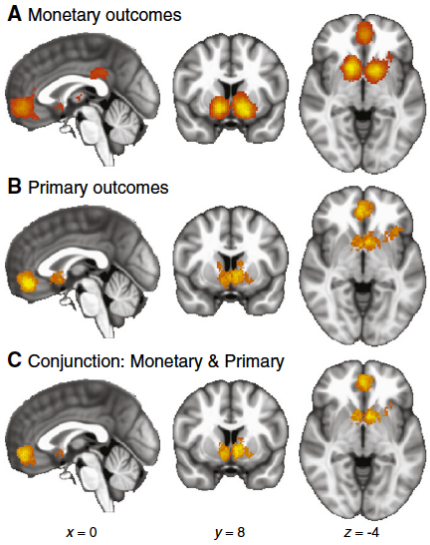


Figure 4.6 Meta-analysis results from whole-brain comparison of activity when monetary and primary outcomes are received. (A) Subjective value of monetary outcomes (B) Subjective value of primary reinforcer outcomes which were immediately consumed. . These included liquids (juice, milkshakes, water, wine), odors, music, attractive faces, and arousing images. Other outcomes which were not unambiguous primary reinforcers were excluded (including pictures of foods and products, guided mental imagery, artwork, humor, signals of social approval, conditioned stimuli, or generic positive feedback messages). (C) The conjunction of areas which activated by monetary and primary outcomes. Source: Bartra et al 2013.

One interesting finding from these authors is that there is substantial overlap in encoding of positive rewards and negative penalties, in caudate and bilateral insula. The authors also analyzed signals at the time of outcome. S=82 studies used money outcomes and S=33 were categorized as "primary outcomes". Primary outcomes are unconditioned stimuli (UCS)— that is, stimuli for which little or no learning is needed to establish valence and value. These include liquids, odors, music, attractive faces, and arousing images. As Figure 4.6 illustrates, there is substantial overlap in regions that encode monetary and primary outcomes. Monetary outcomes are not "primary" reinforcers (or as discussed earlier this chapter, innate or prepared) because children are not born knowing coins and paper currency are valuable. They are either called CS or "secondary" reinforcers. The map of the conjunction of monetary and primary regions is almost the same as the one for primary outcomes— there is nothing very distinctive about encoding of secondary rewards. This is one example of what is meant by "common currency" of value encoding. Learning about the value of secondary outcomes comes to create activity in neural regions which overlap strongly with encoding of primary outcomes.

Of course, the fact that vmPFC is so reliably activated by value in these meta-analyses does not imply that its sole function is to encode value. To see this, it is useful to be reminded of the distinction between "reverse inference" and "forward inference".

A reverse inference is the hypothesis that asserts a high probability P(function|region). In the earliest days of cognitive neuroscience and neuroeconomics this was often just a guess about the function of a region that was identified as active.

A forward inference is the other way around; it is an assertion about P(region|function). If it is easy to generate false positives in identified regions, then a forward inference is better because it is usually predictive and can be specified in advance (even pre-registered).

Reverse inference was widely-criticized in the 2000s as a result of underhypothesizing *ex ante* and bold guessing *ex post*. That is, after seeing what regions were active, a guess was made about the function being performed. Critics were concerned that such guesses were not actually conclusions, but instead were stepping stones to what one might find in later studies— and that they were being overdescribed (or over-accepted by readers or journalists) as solid conclusions. Now the field has come to accept a helpful role for reverse inference, accepting that many such assertions are tentative.

The fact that P(vmPFC|valuation) seems to be high does not imply the reverse inference, that P(valuation|vmPFC) is necessarily high too. vmPFC may be encoding many other computations and be involved in many types of circuitry.

Addressing exactly this possibility, *Nature Neuroscience* (Delgado et al 2016) published an influential and informationally-dense "dialogue" among different pairs of researchers. Each of the three pairs discussed apparently different functions of vmPFC. This engaging dialogue is a few years old, but it still lays bare central questions which are still mostly unanswered, and can be read as a giant, and candid, grant proposal for what kind of research should be done.

1. One function is valuation and value integration. Anterior vmPFC seems to encode experienced utility from consumption experiences. (This evidence was important in the discussion of "wanting versus liking" in Chapter 3.) Lesions to vmPFC also disrupt flexible value-based learning and creates inconsistent, intransitive valuation (we'll discuss this more below).

2. A possibly different vmPFC function is social evaluation of the self and socially-close others (BA 10), and motivated self- and other-description (BA 11).

3. A third and more different function is inhibition of maladaptive affective responses. A lot of these data come from extinction after fear conditioning (i.e., a learned-to-be-fearful stimulus is no longer paired with an aversive UCS such as a shock, which creates extinction that may or not be "reinstated" when the shock appears again). There is also some evidence that similar vmPFC regions are involved in placebo effects and cognitive reappraisal to regulate emotions (see chapter 9).

The dialogue asks challenging and hard-hitting questions about the best data supporting each of the three functional perspectives, the weaknesses of each, and empirical challenges.

Since vmPFC is a large region, it is quite possible that there are subregion specializations (and some evidence of that from single-unit recoding). A different natural instinct (at least for me) was to speculate about whether there is some some higher-level function that could plausibly encompass all three views. For example, there is an obvious link between social information about self and others, and general valuation, if social information is valuable to monitor relationship value and inform social interaction. But Beer and Platt worry that it is hard to dissociate valuation and social inference experimentally (p 1550):

> …we remain deeply suspicious that any test could somehow completely cleave value and affect from any social interaction, thus making it difficult to fully disentangle these perspectives. Whereas social interactions may vary in their degree of value or affect, a social interaction devoid of value or affect seems very unnatural indeed.

Another interesting link is between valuation and affective responses associated with fear and inhibition: In both rodents and humans, there is a well-established "social buffering" which reduces expressed fear of a conditioned stimulus (CS) in the presence of a conspecific animal. To the extent that a fear response is a negative Pavlovian valuation, it seems to be reduced (creating value) by social presence.

A short time earlier, Roy et al (2012) had argued similarly about the range of functions of vmPFC, motivating their important empirical study:

> [vmPFC] serves as a hub that connects systems involved in episodic memory, representation of the affective qualities of sensory events, social cognition, interoceptive signals, and evolutionarily conserved affective physiological and behavioral responses. As such, it plays a unique role in representing conceptual information relevant for survival and in transducing concepts into affective behavioral and physiological responses (p. 148)

These authors tried to figure out how a vmPFC-centered reward circuit is different or similar than other circuits. This is important for social science because it will help us understand how specialized valuation is, and inform the neural basis of the three-approaches dialogue discussion in Delgado et al (2016).

A graphic illustration of their sweeping idea comes courtesy of Neurosynth.[99] Figure 3.11 shows a series of "networks" which are found when keywords are typed into Neurosynth. Factor analysis is then used to divide regions into two orthogonal factors. The loadings on the two factors can be used to compare network overlap (panel (b)).

Default mode and memory networks are entirely factor 1. Emotion and autonomic/endocrine networks are mostly factor 2. Perhaps surprisingly, reward regions show factor loading close to autonomic/endocrine. This is a big clue that reward encoding is not generally highly abstract, as it shares circuitry with 'low-level' biological activity and with

[99] Neurosynth is a website which consists of thousands of published fMRI studies. It uses an automated method to extract peak-activity voxels and the words authors use to describe what that activity is encoding (e.g., reward, emotion, conflict). It is a brilliant help in quickly summarizing the cumulation of many different studies.

emotion. Any social science that treats reward and preference as decoupled from emotion appears to be making a serious mistake (see more in Chapter 9).

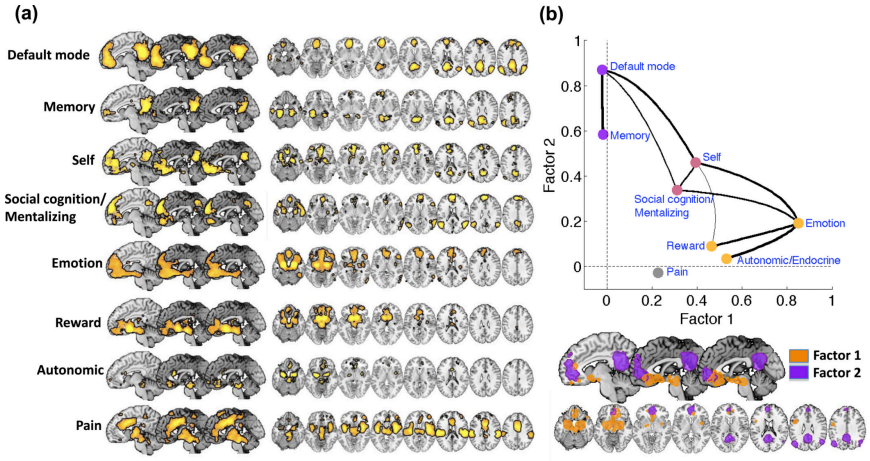


Figure 4.7: Convergence of multiple circuits in vmPFC. (a) Results of an automated reverse inference meta-analysis using Neurosynth. Maps display the reverse-inference probability terms related to vmPFC's functions given observed activation P(term=function|activation). (b) Results of a factor analysis with two factors on the meta-analytic reverse inference patterns. Top panel: Factor loadings associated with each term. Bottom panel: Spatial extent of the regions associated with the two factors (including voxels with loadings in the top 1% of values across the brain). Source: Roy et al (2012).

Two other specialized findings about vmPFC functions are notable. One comes from the mental value of control.

In an experiment, one group of rodents experienced a series of mild shocks that can be turned off if the rodent took an action. These rodents therefore had some control of what shocks they experienced. After measuring what those rats did, a control-deprived group is "yoked" to the same series of shocks. Yoking means that the exact same shock sequence created by the rats in the first group, is then delivered to a yoked-sequence rat. The two rats received the same shocks, but the first rat got to control the shocks while the second rat did not.

Simply having control has many effects: It speeds up learning to escape, slows down how rapidly fear is learned by RL, and reduces serotonin response to threats in new contexts. Importantly, inactivation of vmPFC erases most of these effects. These data suggest that having control is good for adaptation and it requires vmPFC to execute.

The second example is about habits. In goal-directed (model-based) choice, when a food reinforcer is devalued, by feeding to satiation or creating taste-aversion, animals should quit working to get more of the food. In normal goal-directed functioning, animals that are not

overtrained *do* quit working for the food they don't want. However, when there is inactivation of vmPFC they continue working, as if the goal-directed signals about goal value are broken.

vmPFC lesions and rational transitivity

A lot has been learned about vmPFC from studies of lesion patients with damage in that region. What those patients can, and can't do, normally is rare quasi-causal evidence about the normal role of vmPFC in valuation and choice. (There is also direct evidence about neural circuitry of choice inconsistency in which vMPFC is a prominent contributor, Kurtz-David et al 2019).

Kurtz-David, V., Persitz, D., Webb, R. *et al.* The neural computation of inconsistent choice behavior. *Nat Commun* **10**, 1583 (2019). https://doi.org/10.1038/s41467-019-09343-2

There is case report evidence that vmPFC damage can impair decision making. Eslinger and Damasio (1985) describe behavior of one patient with extensive medial orbitofrontal cortex damage from a tumor resection.

> "Deciding where to dine might take hours, as he discussed each restaurant's seating plan, particulars of menu, atmosphere, and management. He would drive to each restaurant to see how busy it was, but even then he could not finally decide which to choose. Purchasing small items required in-depth consideration of brands, prices, and the best method of purchase." (p. 1732)

Yu et al (2018) did a thorough study of patients with damage to vmPFC. This will be used as an exemplar of how lesion studies work. People with "focal" damage to frontal lobes were recruited from two different patient registries at the University of Pennsylvania and McGill University. "Focal" means that damage is limited to compact regions which can be identified by structural MRI. Patient registries find and keep track of patients with various kinds of lesions and disorders.

Disorders were typically caused by aneurysms, strokes, or hemorrhage. Patients were tested from .5 to 17.8 years after damage occurred and had an average age of around 60. This length of time after damage is potentially important because the brain is plastic (recall chapter 1) and people can creative cognitive workarounds after brain damage to maintain task performance (just as a right-hander who breaks their right hand can learn to manage with the left hand over time).

The subjects were N=13 people with ventromedial (VM) damage, and another N=10 with frontal damage outside the ventromedial area (called frontal controls, FC). There is also a second control group of N=12 "healthy controls", HC) who are roughly matched on a few demographic variables (usually age, gender, and education). The idea is that if the VM patients perform abnormally on a task, that abnormality is evidence that the VM is *necessary* for doing the task well (recall Chapter 1 methods).

Each trial is a choice of one of two objects from a category (e.g., paintings). The main dependent variable is how often "transitivity" is violated. Transitivity the property that if $A \succ B$, $B \succ C$ then $A \succ C$ (where $A \succ B$ denotes that A is preferred and chosen over B). Transitivity is an

important property for choices to obey because the number relation "bigger than" is transitive. If choice preferences are not transitive then the brain must not be using stable subjective numerical values (or ranks) to determine choice. (The featured study by Li et al in Chapter 6 gives more detail of this type of rationality test.)

Rates of intransitivity were 9.9% for VM patients, compared to 9.1% and 5.7% for FC and HC patients. The VM-HC gap, while small in magnitude, is significant and replicates the findings of two previous studies.

This low rate of transitivity violations is surprising: Given all the evidence of vmPFC in valuation and choice, and the anecdotal stories about choice-paralyzed patients like EVR, how can those valuation processes be so robust to serious brain damage? Most patients did not have damage throughout all of vmPFC, so it may be that enough adequate function remains in undamaged tissue. Or it could be that sensory or memory systems execute an effective backup, or other parieto-cortical circuitry regions take over.

Shifting Control between Model-Free and Model-Based Systems

Model-free valuation requires substantial experience, is computationally easy, and very likely 'takes over' in a stationary environment where repeating the best action over and over is a good idea (that's habit).[100] However, by definition model-free choice will make mistakes when the underlying reward structure changes.

Given remarkable human flexibility (compared to other species), there must be *some* neural system for arbitrating between the model free and model-based systems, when the behavior they recommend conflicts. In particular, it would be adaptively useful, for model-based valuation to seize control, either when the habitual model-free choices has made large mistakes, or when a dramatic environmental shift is cognitively evident (before the model-free trigger is pulled).

Daw et al (2005Z) were the first to hypothesize an arbitration system that weighed quality of choices by the model free and model based systems, shifting control back and forth when useful.[101]

Keremati et al (2011) present a model in which the model-based system compares reward distributions of different choices and computes a value of perfect information (VPI). The VPI is the increase in the maximal action value which results from deliberation that sharpens the subjective reward distribution of different actions (compared to the no-deliberation current maximum). When choices are ongoing, the model-directed system will deliberate if the VPI is greater than the opportunity cost of choosing (based on the currently-best action). Habit results

[100] A lesion study showed that overtrained rats, exhibiting habits, shifted to goal-directed choice after lesions of prefrontal cortex (Coutureau and Killcross, 2003).

[101] This idea is similar to "rule learning" used in behavioral game theory, in which different rules are weighted by their achieved payoffs (Stahl, 1996).

when the value distributions are sufficiently well-estimated that the VPI is low so that it is optimal—considering deliberation time—to just make the same choice over and over.

(START) FEATURED STUDY: Arbitration between model-free and model-based systems

Lee, Shimojo, and O'Doherty (2014) made an important advance in mental arbitration, by comparing six different models and finding a distinct winner. They also observed areas that encode value, arbitration, and choice (assuming the brain is computing values using the arbitration system that fit choice the best).

They used the two-stage decision tree task developed in Glascher et al (2010) and shown in Figure 4.8. A state consists of a point at which a decision must be made (i.e., a decision node, or a choice set). At the initial state there are two choices (represented by fractal images), corresponding to choosing Left or Right in the tree. Each choice transitions probabilistically to one of four second-stage states. (In RL notation, this probability is T(s,a,s')). Each of those four states also consist of two Left-Right choices represented by different fractals. Those second-stage choices lead probabilistically to one of two reward outcomes, which ends a trial. While the tree is "bushy" visually, it is the minimally-complicated task in which to observe a type of error associated with mispredicting the transitions between states.

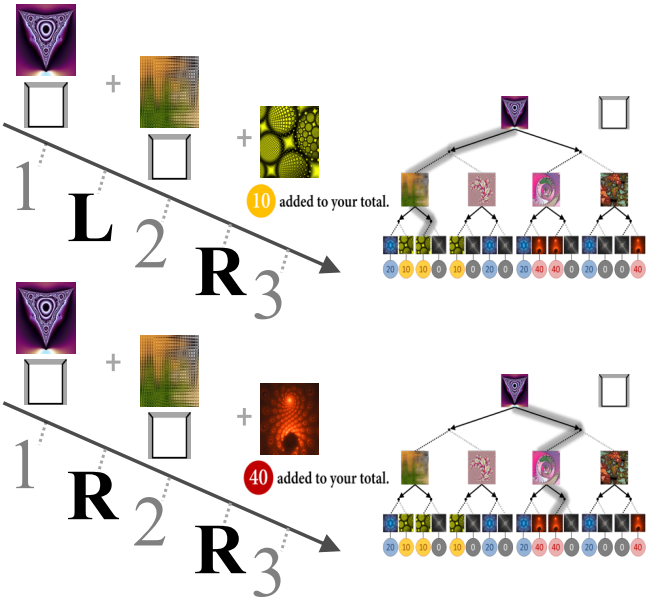


Figure 4.8. This figure depicts a decision tree designed to identify arbitration between model-free and model-based systems. Participants move from one state to the other with a state-transition probability p following a binary choice (of left or right). The temporal sequence of pictured fractals (left) correspond to paths through the decision trees shown on the right. In the top example, a person chooses to go left from the first fractal. Then a random event chooses a transition to the left one of the two fractals. The active participant then chooses Left and a random event chooses the fractal #3 (from the left). This adds +10 to rewards. Source: Lee et al 2014.

If the state *s'* occurs after choosing action *a* in earlier state *s*, the state prediction error (SPE) is $\delta_{SPE} = 1\text{-}T(s,a,s')$. This is like a reward prediction error except that since the transition to s' does not lead to a deterministic known reward, it is just a way to learn about the state transition probabilities. They assume that this error is used to update the estimated transition probability according to

$$T_{t+1}(s,a,s') = T_t(s,a,s') + \eta\delta_{SPE}(t)$$

where η is a state learning rate. They also assume the model-based system makes a "forward" value calculation of the form

$$Q_{FWD}(s,a) = \sum_{s'} T(s,a,s')[\, r(s') + \operatorname{argmax}_{a'} Q_{FWD}(s',a') \,]$$

Note that this forward-looking dynamic programming approach assumes that in state *s*, people anticipate all the states *s'* that can be transitions from the current state *s*, and simulate or recall the optimal choices *a'(s')* in those imagined future states *s'*.

The state prediction error is used for updating T(s,a,s') and hence changes the subjective value $Q_{FWD}(s,a)$. The model-based system is trying to compute or learn the value of first-stage state-action pairs (s,a)— i.e., what are the best choices from the set "left or right" available at each of the two stages?

The Lee et al. version of this two-stage task also includes two reward structures. In a one-color "specific" condition, subjects know that they only earn the reward if one of three colored outcomes occurs (they know in advance what the color is). In an any-color "flexible" condition the color doesn't matter; all rewards are earned (that is the condition shown in Figure 3.10).

In the one-color trials, the model-based approach can use information about which choices in the first and second states/stages are most likely to lead to the right color. The model-free approach does not track that information. Suppose the rewarding color is yellow. If a second-stage choice has led to high reward in the past, but will give no yellow outcomes, the

model-free system will recommend that choice anyway. Thus, the arbitrator should tilt away from model-free and toward the model-based prediction in the one-color trials. To account for the fact that model-free is computationally easier, their approach builds in a bias toward model-free, when computed Q-values are equal.

The arbitrator is like an (American-style) football head coach who listens to two assistant coaches for advice about what kind of play to run. The first assistant just keeps a mental scorecard of the average number of yards gained by either running or passing—those are the action choices-- that day. This kind of simple statistic is easy to compute; indeed, yardage averages are often periodically posted on TV. This coach is using a model-free approach, tracking the average reward of actions, which are *not conditioned* on changes in the game score or in what players are on the field (i.e., states).

The second assistant uses sports analytics. She keeps track of reward histories for actions conditioned in states— field position, time remaining on the clock, which quarterback is at the helm, what injuries have sidelined players, and so on. She also forecasts likely yardage gains in new states for which no recent reward history is available in the current game, using a model-based procedure (such as results from previous games in similar states).

Of course, the first assistant is quick to offer advice. The second assistant is slower because she has more numbers to crunch and more imagining to do. Waiting for better advice is costly because time is valuable in football. In a fast-paced game, it is reasonable for the head coach to just take the first assistant's advice unless the first assistant's advice produces a string of mistakes, or when that advice is sharply contradicted by the second assistant. In the latter case, the head coach might think about which kind of advice has been more helpful earlier in the game— the reliability weight.

Lee et al.'s approach has three computational levels: ongoing model-based and model-free calculation of state-action and action values; measures of reliability of model-based and model-free as ways of generating reward; and an arbitrator which weights the model-based and model-free choices, given their reliabilities, to determine a specific choice in one trial.

Figure 4.9 shows a schematic of what is being computed. Calculating the Q(s,a) values can be done using methods that are now familiar (the top "Reinforcement learning" boxes).[102] What remains is to estimate numerical reliability, and combine two potential choice values incorporating reliability. For that last step, Lee et al use a weighted-average decision rule, in which the reliabilities weight the predicted Q-values ("integrated value computation" in Figure 3.11).

[102] The MF specification is a SARSA rule (named because it computes Q(s,a,r,s',a')). The MB specification is given in (3.4), with the addition of "backward planning" which substitutes the correct r(s') into (3.4) when the person is alerted that the one-color condition is in effect.

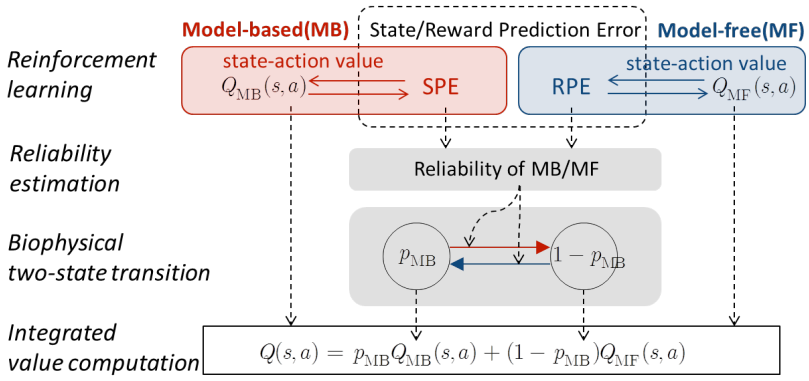


Figure 4.9: This graphic shows the computational elements of arbitration between model-free MF and model-based MB arbitration system. (Top) Two iteratively-coputed Q-values QMB(s,a) and QMF(s,a) are computed from choices and rewards, using Bayesian and Pearce-Hall associations respectively. By assumption, the model reliabilities are Bayesian updates from a Dirchlet prior (which simplifies probabilities as relative counts) on prediction errors categorized as positive, negative or near-zero. A high reliability is defined as having a p_{MB} of near-zero error. The hypothesized arbitration mechanism creates an integrated value computation which is the reliability-weighted average of MF and MB values. Source: Lee et al (2014)

They consider several methods for estimating reliabilities. The Bayesian method divides RPE prediction errors into three categories: positive, negative, and near-zero.[103] Using a Dirichlet prior over the probabilities θ_c of these three categories enables a simple Bayesian procedure that updates the prior by counting new observations. For each of the three categories c, the inverse of the dispersion index (or Fano factor), $\chi_c=E(\theta_c)/\sigma^2(\theta_c)$ is computed. The model reliability is defined as $\chi_{MB}= \chi_0/\Sigma_c\chi_c$. Intuitively, a higher mean $P(\theta_0)$ and lower uncertainty both indicate higher reliability.

For MF reliability they also consider a non-Bayesian method: a quantity Ω tracks the history of absolute RPEs, updated by $\Omega_{t+1}= \Omega_t+\eta(|RPE(t)| - \Omega_t)$. Reliability is defined by $\chi_{MF}=(RPE_{max}- \Omega)/RPE_{max}$.

Finally, the weight p_{MB} , which represents the extent of model-based control, gradually changes according to (a specification taken from biophysical neuronal models (Dayan and Abbott, 2011)

$$p_{MB}(t+1)= p_{MB}(t) + \alpha(\chi_{MF})(1- p_{MB}) - \beta(\chi_{MB})p_{MB} \qquad (3.5)$$

[103] The near-zero bin is defined by a threshold free parameter ω; if RPE< ω it is set to zero.

$$\text{where} \quad \alpha(\chi_{MF}) = A\alpha/(1+\exp(B\alpha\ \chi_{MF})) \quad \text{and} \quad \beta(\chi_{MB}) = A\beta/(1+\exp(B\beta\ \chi_{MB})) \qquad (3.6)$$

They compare several combinations of Bayesian and non-Bayesian specifications, with the dynamical transitions of control as in (3.5) and no such transitions. Bayesian updating for MB and non-Bayesian for MF, with dynamical transition—called "mixedArb"-- fits slightly better and becomes the focus of fMRI analysis.[104]

The "mixedArb" arbitrator model captures key features of the data on how people actually behave. Choices consistent with MF have slower RTs than MB-consistent choices for most subjects. The model is well-calibrated in a statistical sense: If you bin all trials in which the model predicts Right will be chosen with probability around .8, for example, the relative frequency of Right choices across trials in that bin is indeed around 80%. The model also predicts choice switching in a sensible way: When P_MB<.5, the MF system is largely in control, and choices usually do not switch from trial to trial. When P_MB>.5, there is more switching. Finally, the MB and MF reliabilities, as encoded in the best-fitting specification activate brain areas (as we'll see very shortly). However, the same kind of reliabilities, as measured using other specifications, do not activate *any areas* with high significance. Thus, we can think of the better mixedArb specification as both fitting behavior more accurately, and "fitting the brain" more accurately, in the sense that *some* activity appears which is correlated with the mixedArb reliabilities.

What else happens in the brain? Figure 4.10 shows that value signals associated with MB and MF Q-values, and the difference in the arbitrated (reliably-weighted) values of the chosen and unchosen actions, are encoded in typical value areas. For example, the third slice from the left in the bottom row of Figure 3.14 highlights activity in posterior L putamen (in blue) in response to MF value only. Putamen is an important area because activity there is seen in overlearning, persistence of habit after devaluation, and closer DTI connectivity with premotor cortex for people who make more habit-related action slips (Wunderlich et al., 2012; Tricomi et al., 2009; de Wit et al., 2012).

Encoding of arbitrator weights was found in bilateral inferior lateral PFC (=ilPFC) and right fronto-polar cortex (rFPC).[105] The next step is to see if there is a sensible link between value computation and reliability. They used the bilateral ilPFC and rFPC regions as physiological seeds in a PPI analysis, using the computed value P_{MB} as the psychological moderator. The

[104] Interestingly, the various arbitrator specifications do not fit *that* much better than a simple model with only MB (the difference is only 5% in log likelihood; see Lee et al. (2014) Table S1). An important next step is to design tasks in which different specifications are *ex ante* expected to be easier to distinguish statistically.

[105]This region is also activated when choosing gambles with epistemic "ambiguity", in the sense of missing important information or knowing little about the domain of the gamble's risk (Hsu et al 2005 Science TO DO ADD CITE).

analysis turns up regions that have activity which correlated with seed-region activity in a way that varies with P_{MB}.

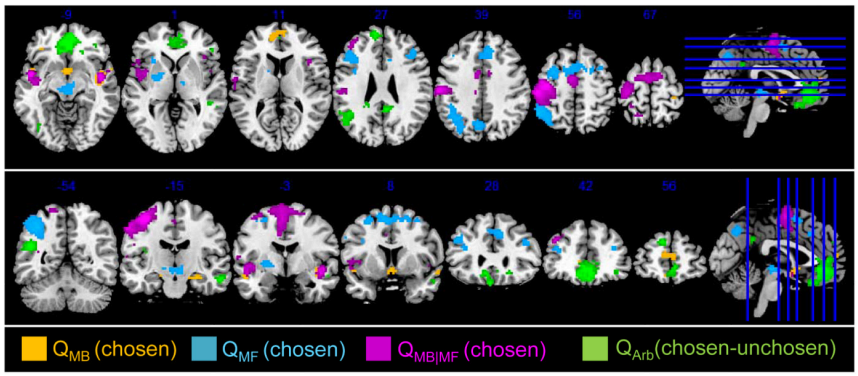


Figure 4.10. Neural activity encoding model-based and model-free value signals. Q_{MB} is the value of the action 'chosen by' the model-based system (i.e., the choice with the highest value) . Q_{MF} refers to the chosen value of the model-free system. Areas described as $Q_{MB|MF}$ respond to chosen values commonly for both systems. Q_{Arb} is encoding of the chosen minus unchosen value signals, in which the value signals are a weighted combination of model-based and model-free values determined by the output of the arbitrator (P_{MB}). Source: Lee et al (2014)

They find a *negative* link between ilPFC and rFPC and putamen depending on P_{MB}. This is what you would expect to see if, when P_{MB} is high, the arbitrator is favoring the MB choice and suppressing activity related to MF input (which appears to be in putamen, as it was established to encode Q_{MB} in the earlier analysis).

Note that two other studies found activity in right lateral PFC, close to the ilPFC and rPFC regions shown in Figure 4.10. Those studies found that rPFC encoded decision confidence and metacognition (DeMartino et al 2013; Baird et al 2013). Hence, this small number of studies suggest these regions may be involved in general types of higher-order uncertainty computations. A question often asked in modern decision neuroscience is whether the MB and MF systems *compete* or *cooperate*. Probably they do both in different ways. For example, in the arbitrator approach, the two systems clearly *compete,* because the more reliable system gets weighted heavily in the arbitrated Q-value computation. However, note that there is also an implicit division of labor between the systems, since the model-free system is fast and approximately accurate under stable learning,[106] while the model-based system is superior in more demanding tasks (the Lee et al one-color condition) and under conditions of changes in reward. In a stationary two-stage task, with no shifts from one-color to any6-color reward, the arbitrator model generates a P_{MB} which starts around .70 and exponentially declines to around .10 after 160

trials. Model-free habit takes over because, from the two-system point of view, it *should* take over.

Furthermore, the MB system's choices also provide input to RPE learning that shapes MF valuation.[107] Indeed, because it is trying to learn structure and make optimal choices, the MB system is, indirectly, providing the ideal raw material (RPEs) for the MF system to learn, and take over control when rewards are stable.

For example, suppose in the two-stage tree task that Left is a good choice and has high subjective valuation under both MB and MF (i.e. Q(s,Left)>Q(s,Right)). Now suppose the reward structure changes so the reward to Right is suddenly greater than to Left. The best way to teach the MF system how to change its subjective value is to sample extensively from Right, and then to make the best second-stage choices after Right is chosen in the first stage. Those changes are exactly what the MB system will do. When MB choice takes control after the reliability of MF plunges, its choices will act as if they are optimal given modeled-structure, but also have the ideal long-run effect of helping simple MF get back on its feet and take over when ready.
END FEATURED STUDY

Section Summary

Model-directed choice is an effortful, deliberative mode way of choosing which constructs or assembles an estimate of the distribution of reward value of outcome O after action A. In the heyday of behaviorism, "mentalist" constructs were eschewed in favor of the hope of explaining everything from a single set of S-R building blocks. But even in that heyday, there was clear evidence from Tolman and his Tolmaniac believers that animals learn reward-related features of their physical world and available actions.

A workhorse task to study model-free S-R and model-directed A-O was created by Daw et al (2011). The key to this two-stage task is that an action which is rewarding can also convey information about how an unchosen action is even more rewarding. (A similar idea appeared in Camerer and Ho, 1999, in the form of attention to foregone payoffs, to hybridize reinforcement and belief learning in games.)

The two-stage task generates robust evidence that even educated adults use a mixture of model-free and model-based learning. Cognitive load, youth, and various disorders are associated with less effortful model-free learning.

Just as the two-stage task became popular, it became obvious that there was a likely speed-accuracy tradeoff between the two systems. Understanding how control is shifted back and forth between the systems is now an active area of research, as illustrated by the Lee et al (2014) arbitration data. This is also a central question in modelling habits, as discussed in chapter 1.

---

[107] The opposite is, of course, true also: The MF system makes choices that generate state prediction errors SPEs as well.

Hierarchical Reinforcement Learning

Chapter 3 described in detail the "two-stage choice task" (Daw et al 2011) used to contrast model-free (MF) and model-based (MB) valuation systems. Two stages are useful for that purpose because they create the minimal decision tree that can separate model-free and model-based choice (one stage can't do it).

However, in the two-stage problem with two branches at each stage, it is easy to learn the value of all the paths through the tree because there are only four paths. The simplicity of that example inadvertently hides the fact that in general, simple versions of reinforcement learning may do much more poorly in complex sequential domains. In more complex domains, reinforcement learning learns too slowly—and more slowly than people do—because there is too much to learn (Botvinick, Niv Barto 2009) from trial-and-error.

Consider making coffee in the morning: Doing so requires getting a cup, putting coffee grounds into a coffeemaker, checking the water level and adding water if needed, pressing the right button, pouring the coffee, adding milk, and so on. If each of these steps is treated as a branch in a tree, and some steps are equally effective in different orders, the tree quickly grows "bushy". Learning the best path through the tree through simple reinforcement learning is difficult. In computer science and RL this difficulty is called the "scaling problem" (or the problem of "option discovery").

Like many combinatorial explosions, it is easy to see how bad the reinforcement learning scaling problem is by using big numbers. Botvinick et al (2015) scare us with an example:

> "As an illustration, imagine a video game in which the goal is to find a treasure by making a series of twenty choices between pairs of distinctively decorated doors, with each door leading to a unique pair of successors [new doors]. Assuming one played this non-stop, choosing a new series of doors every waking minute, discovering the treasure would take an average of one and a half years." (p. 72)

A sensible solution to this combinatorial explosion is that a series of low-level actions are somehow "chunked" into a sequence based on abstract conceptual representation (as suggested, for example, by Lashley, 1951 and later emphasized by Herbert Simon).

The idea of chunking is everywhere in psychology and neuroscience. For example, most people cannot remember a string of 10 digits. But they can remember a 10-digit American telephone number, because the first three digits are a chunked area code, and the next seven digits are usually written as two chunks of three and four digits. If the area code is well chunked, remembering the 10-digit code is the same as remembering 8 units. In the coffee-making example, the sequence of steps is chunked into a subroutine "making coffee". When the subroutine is chosen, each step is automatically carried out.

Recall that we discussed chunking indirectly in chapter 3, with reference to how habits are often motor sequences that are chunked together, and hence can be difficult to break apart once initiated. Each action in the sequence, when executed, is like a Pavlovian stimulus which

triggers the next step in the habit sequence.[108] In other areas of cognitive psychology, chunking is central. Two examples are systems introduced in the earliest days of cognitive science (called SOAR (Lehman et al 1996) and ACT-R production systems (Anderson 2004) architectures).

In some cases, such as grooming sequences in rodents, (Aldridge and Berridge, 1998), sequence execution seems to be somewhat innately prepared genetically; it does not require much require learning. For complex human actions, however, learning of some kind is needed, especially when the action sequences are culturally influenced.

Dezfouli and Balleine (2012, 2013) used the chunked habit-sequence concept to make progress on understanding a simple form of hierarchical reinforcement learning (HRL). HRL is a special kind of model-based learning in which what is learned is the value of sequences that form a hierarchy. The simplest example of such a hierarchy is a decision tree, in which the first actions in a sequence are at the top of a temporal hierarchy.

The authors used the familiar two-stage task. Recall that in that task, people can choose two left or right actions (labeled A1 and A2 here) at each of two stages. Let's label the second-stage actions B1 and B2. Thus, there are four possible sequences of two-stage actions (A1B1, A1B2, A2B1 and A2B2).

The authors contrasted two decision architectures. In a hierarchical architecture, once the habitual sequence is initiated, it continues through both stage of choice until reward. Habit cannot be interrupted by further deliberation. The opposite is a "flat" structure in which the habitually-prescribed case can be interrupted by a model-based override in the second stage.

The key test of whether habit is uninterrupted in the second stage, or sometimes interrupted, is this: Suppose a person chose the sequence A1B2 on the previous trial and earned a reward. Then a new trial starts. Suppose the first choice was made rapidly (an indicator of habit), and the choice was A1. The A1 choice matched the first-stage choice on the previous trial. In the hierarchical structure, choosing A1 initiates the sequence A1B2, so that the subject is very likely to choose B2 again (i.e., to "stay" with the same choice that rewarded in the previous trial). In the flat structure B2 may be interrupted by a model based calculation to choose B1 instead. However, persistence of choice of the habit-initiated B2 is what the authors observed (their Figure 8).

The next question, a challenging one, is how much longer sequences are carved into chunks. A sensible start is to recognize that many decision sequences have a natural physical or sensory "bottleneck". A bottleneck is a state, or small number of states, that all sequences must pass through. Think of a tollbooth on a busy highway. There are a lot of driving paths before the toolbooth, and after, but all paths must pass through the tollbooth bottleneck.

[108] A marker of this type of chained-sequence habituation is when monkeys, performing a motor sequence for reward, continue the entire sequence even if the reward is delivered prematurely (e.g. Matsumoto et al 1999).

Bottlenecks create natural versions of what computational neuroscientists call “subgoals”. A subgoal is an earlier goal state that must be passed through to continue making progress toward and end goal. If you don’t pay the toll going through the one tollbooth on the New Jersey turnpike (reaching the subgoal) you do not get to see your cousins in Somerville (the goal).

A good component option is a path from one subgoal to the next subgoal, ending at the final goal state. Model-based valuation can value these component options but if they are restricted in number then RL can learn them quickly too. Ideal paths between subgoals are more learnable by RL if the combinatorial explosion of possible paths is choked off by bottlenecks.

Botvinick et al (2105) note that bottlenecks are like edges in perception. Efficient coding of complex stimuli is thought to consist of distinctions and categorical equivalences that both learn from, and therefore reproduce, natural ecological statistics of co-occcurence and redundancy. If you only see Buddy walking his dog Happy around the neighborhood, it is efficient to perceive and remember them as a single “boy-and-his-dog” unit rather than waste perceptual capacity on two separate perceptions. The hypothesis that a mental system is doing this boy-dog binding predicts that a neighbor will be surprised by Buddy walking along without his dog, or when Happy escapes and runs along off the leash.

Hierarchical reinforcement learning, beyond the two-stage case, is well illustrated by a paradigmatic spatial navigation problem introduced by Sutton et al (1999). It is called the “rooms” problem. The rooms problem is a 2D grid of boxes. Each box represents a location in each of four rooms in a building (see Figure 3.15a). As you can see, a room is a collection of adjacent boxes which is only connected to an adjacent room by a “door” box (a foyer, really), permitting passage from one room to another. A person starts at one of the boxes (marked S) and must reach a specified goal state in another room (marked G). By the way, computer scientists love canonical test problems like the rooms problem. The problems that become popular are somewhat lifelike, not too easy, and can be complicated in many dimensions. They are like “model systems” in biology or “toy models” used in economics (such as Bertrand competition).

The baseline learning mechanism we will start with is reinforcement learning over “primitive actions”. Primitive actions are movements from one box to one of the eight adjacent boxes next to it. Boxes are states. In a familiar actor-critic architecture, the actor chooses a policy $\pi(s)$ which is a transition function from all of the possible boxes to each box that is adjacent. The critic learns the value $V(s)$ of being at a particular box-state.

This algorithm begins learning with no information. It is rewarded when it stumbles upon the goal state. Temporal difference learning propagates learning the value of the goal state, when it is reached, backwards to states along a good path toward the goal. For example, when the algorithm first reaches the goal state, there is a reward prediction error which updates the value $V(s)$ of the state representing the box just before the goal state was reached.

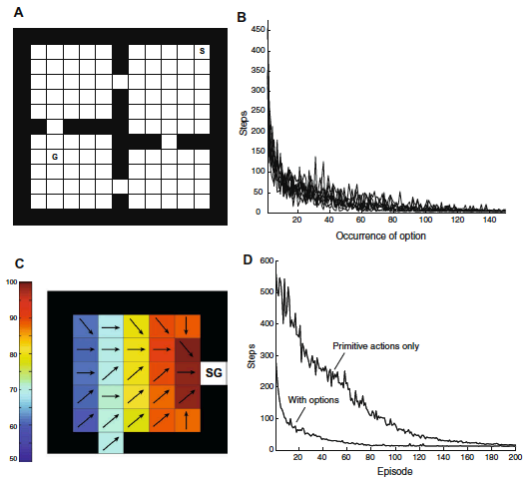


Figure 4.11. The "rooms" problem". Agents move from one box to another in steps. (A) shows collections of four adjacent boxes which are rooms. Each of the four rooms has two boxes which are "doors", connecting rooms. (B) Learning curves for the number of steps needed to go from start box S to goal box G. (C) This describes activity in boxes only in the upper-left room, (important) when the S and G locations are switched so that an agent desires to go through the rightmost door. The box "SG" is a "subgoal" which transitions from the upper left room door to the upper right rooms. The heatmap for each box expresses the learned value of being at each box-state. The darker colors near the SG box are the model's way of saying that when the boxes closer to the SG subgoal have higher value (reward is 'closer'). (D) Learning surves for how many steps are required to reach G. When subgoal options are added—that is, the agent learns value of bottleneck door subgoal options—learning is much more rapid.

Figure 4.11shows some statistics on how fast an actor-critic model can learn to get from the box marked G to the goal state marked S. In the first ten trials or so, it takes 50-400 steps before a person finds a way through two doorways to the goal. (The optimum is 11).

Now suppose the person can learn chunked "options". Options are box-to-box transitions supplemented by the ability to learn the value at a doorway subgoal, rather than only learning from reaching the eventual goal. The subgoal values are still learned by TD; that is, the subgoal nearest to the reached goal gets updated positively.

In this process, when the option to head toward a subgoal is chosen, a smaller number of steps is executed to get there. If a learner is in the upper left room in Figure 4.11a, the subgoal is the door located in the third row, at the sixth box on the right (jutting out from the other five-box rows). Figure 4.11ctakes that upper left room and magnifies it. In Figure 4.11c, the frequencies through various boxes in the room to the doorway subgoal are shown by arrows (indicating the most likely *direction* of movement out of one box) and darker colors show the relative frequency of the most common direction. The darkest colors show that if you are in boxes next or adjacent

to the subgoal box marked SG, the relative frequency of moving toward it is very high. (That is also what a bottleneck looks like.)

Figure 4.11dshows the number of steps taken to reach the final goal. The learning curve is much steeper when subgoal learning (denoted "with options") is allowed. After about 80 episodes, the learning with subgoal options is close to the optimum of 11 steps. Without options full learning takes hundreds of trials.

Thus, hierarchical reinforcement learning which includes specialized learning about "shortcut" doorway options is clearly superior. It also degrades less badly when there is a 'memory loss' requiring the learner to start over (which is not shown in these Figures). However, there is a subtle limit in hierarchical reinforcement learning: When the room structure changes, it can persist using an old decision policy that is no longer optimal, and cannot be easily unlearned. In cognitive psychology, this phenomenon is generally called "negative transfer" of learning[109]. The authors give two simple examples of negative transfer in the rooms problem.

One intriguing recent idea is that discovery of subgoals is intrinsically rewarding—that is, there is a mental reward from having the subgoal insight, beyond the increase in extrinsic reward from completing the sequence of subgoals to earn an eventual reward. Such an intrinsic reward would motivate a specialized search for subgoals, and perhaps enhance memory of them.[110] One possibility is that subgoals have unusual salience, and recreating occurrence of salient events is intrinsically motivating. (This is undoubtedly true outside these highly stylized examples too, when people find shortcuts or "hacks".) Under these reward-enhancing conditions, Singh et al (2005) demonstrate computationally that hierarchical skills will arise from stepwise development. The fact that dopaminergic neurons which respond to reward prediction error also respond to salient events (independent of reward, e.g., Bunzeck and Duzel, 2006)) is a key piece of evidence for this view. Reaching subgoals may be generically salient in a way that activates the first-component of dopaminergic firing in the two-component Schultz (2019) account from the last chapter.

There is some neural evidence supporting the hypothesis that hierarchical reinforcement learning is common in human cognition. High-level sequential options seem to be encoded as "task sets" in DLPFC, which then selects among these stimulus-response pathways, and which are then implemented by activity outside prefrontal cortex (e.g. Miller and Cohen, 2001). Other frontal areas, particularly pre-supplementary motor area (pre-SMA) and SMA, are thought to

[109] The concept of negative transfer could prove useful for behavioral economics. As these examples illustrate, learning algorithms that work well in general can backfire when faced with changes in environmental structure. The very properties that make learning rules good under stationarity—rapid convergence, and "lock-in" on good solutions— can make it hard to unlearn. This general concept could have application for some kinds of decision making in the human lifecycle, and for organizational routines resulting in "dinosaur" business firms.

[110] In human social settings, it is also possible that subgoal discovery which can be shared in a group creates reward in the form of social status or prestige. In navigating uncharted territory, for example, a subgoal discovery could be finding a narrow waterway that connect larger lakes or oceans (such as the Straits of Magellan in Chile, which shaved hundreds of miles off of Drake's Passage around the ocean-exposed Cape Horn).

encode action sequences, as indicated by single-unit recording of neural firing (e.g. Shima and Tanji 2000).

Dorsolateral striatum (DLS) is a candidate region for encoding option-specific policies. All the frontal regions just mentioned (DLPFC, pre-SMA, SMA), plus primary motor cortex (PMD), have projections to DLS. Rougier et al. (2005) show computationally precisely how this frontal projection can select among different S-R pathways. DLS neurons are also sensitive to the environmental state, which is necessary for switching among options. For example, different DLS neurons fire depending on whether a rat's primitive grooming action is performed with no other actions, or as part of an extended grooming sequence (Aldridge and Berridge 1998).

Diuk et al (2014) looked for prediction errors associated with learning values in a hierarchical two-stage task. In the first stage people picked one of two "casinos", each of which leads to a second-stage choice from a set of four "slots (with different quadruples for each casino). They found different RPEs for each of the two hierarchical stages in a common VStr region (although the RPEs are statistically distinct)

Finally, OFC is a candidate region for representing the value of the state associated with an option subgoal. That value indicates the likely reward associated with the subgoal (like a continuation value in dynamic programming), and that value must be linked with the subgoal identity in order to guide choice among option sequences. Any region that is a candidate for representing the subgoal value effectively for action must very probably be linked functionally with ventral striatum (VStr), which encodes action values, and DLPFC, which represents the steps in an action sequence. OFC has strong connections to both regions. Furthermore, OFC neurons are sensitive to shifts in task set (e.g., O'Doherty, Critchley, Deichmann, & Dolan, 2003).) and change encoded event values when strategies change (Schoenbaum,Chiba, and Gallagher (1999))

Section summary: A long-standing challenge for reinforcement learning is that for complex problem spaces, learning by trial-and-error S-R-O feedback is too slow to match the pace of human learning (or to be useful in building artificial systems). A student learning to play chess does not learn by trying out zillions of sequences of moves and assigning win and loss values to each sequence. Hierarchical reinforcement learning refers to simple RL principles applied to model-based structures which have robust, flexible properties to help learning work faster. The "rooms" problem we discussed is a good example: Many lifelike problems have small numbers of action choices at bottlenecks; these bottlenecks create an architecture of sequences that are all required to reach a (smaller number) of subgoals. It is then faster and easier to learn subgoal reward values (akin to continuation values in dynamic programming). Research in hierarchical RL is proceeding very rapidly in neuroeconomics.

Conclusion

This chapter adds to the discussion of the last Chapter 3 about general neural mechanisms that produce simple kinds of valuation and choice. This chapter focused on model-based valuation, in which a more elaborate representation connecting choice environments, actions and sequences of actions, and subsequent outcomes, guide choice. Decision trees are a good generic and familiar examples of such models.

The coexistence of model-free (MF) and model-based (MB) valuation is not consistent with the simpler assumption of stable preferences, reliably revealed by previous choice and predictive of future choices. Model-based valuations might stabilize, and could then be thought of as stable preferences if associated rewards are not too variable. Even then, preferences will obviously change through the biological lifecycle or ontogeny, when social circumstances change, and when entirely new products or services appear.

A learning process that continuously learns subjective value is therefore the appropriate primitive construct. The learning models are also fertile because they contain many ideas about the factors which may or may not lead to choices that are consistent with stable preferences (in the translated form of stable "reward predictions").

Consider a professor who eats lunch every working day at the small postage-stamp campus of Caltech. She will soon come to learn subjective values of preferred meals at the only four major locations for lunch. The subjective values are mostly fixed in the medium run of a year or two. There is plenty of neural evidence that for small choice sets that are familiar, or become familiar in the time scale of several days worth of training, stable subjective value signals are actually encoded in the brain (e.g. Padoa-Schioppa and Assad, 2008). This neural evidence probably generalizes to what our professsor's brain encodes.

After a year or so at Caltech, a new professor using model-free choice might appear, from observed choice, to be maximizing subjective value given stable preferences. But she will do so because of how RL and reward stability neurally implemented a sequence of choices that reach a plateau of stable subjective valuation and subsequent choice. Her choices are consistent with stable revealed preferences but they were produced by a particular neural procedure.

Note that conventional assumptions of stable preferences and optimization in rational choice social science are sometimes coupled with assumptions about foresight or self-awareness. But model-free choice has no such self-awareness (that's what we mean by saying it is "cognitively inflexible"). That is why, in chapter 2, we introduced the important distincton between implicit processing (of which model-free learning is often example) and explicit conscious awareness of the reasons for choice.

In the professor's lunch example, the difference between MF and MB becomes evident when a new place for lunch opens. Suppose the Red Door café opens after the pandemic closure. In narrow MF mode the professor will ignore the new Red Door options and get her "usual". In MB mode she will have some modelled belief, perhaps informed by generalization from experienced options psychologically similar to new ones described on a menu (even though she hasn't experienced then) or by other sources of information about the new options. The MB-

driven professor, especially if she is adding an "exploration bonus" to new options (as in optimal bandit choice), will try out the new Red Door options.

An example in field data in which MF and MB models can predict interesting differences is the effect of changes in hours at which liquor stores are open. In a model-based rational addiction framework, optimizing alcoholics maximize the amount they drink, with foresight about the effect of current consumption on future preferences. The foresight is what makes their choice "rational" in the special sense of accurately forward-looking.

Suppose a policy change is made so that stores have shorter night time or weekend hours. Reduced hours are a common change in many areas and countries, typically influenced by religion or governmental paternalism. Rational addicts will just stock up when the stores are open, fully anticipating their reduced opening hours. Non-addicts will too. In rational addiction models there is, therefore, no predicted drop in total liquor sales,

However, most studies of this type of policy change indicate that sales *do* drop. Let's stipulate this as a fact in the discussion that follows.

Neuroeconomic models "explain" this fact as a consequence of at least some behavior being guided by MF habit mode. In habit mode, people only buy liquor when physical state-cues associated with regular purchase-- such as seeing an open store, or thinking about weekend plans on Friday-- create instant "demand". No such cues means there is no demand. This is a radical claim from the point of view of rational choice social science because in this story there are no stable preferences of the usual sort—there is, instead, a roughly automatic state-dependence of action based on physical state cues.

However, the word "explain" is in quotes in the paragraph above, because this type of neuroeconomics theory is not a satisfactory explanation until it is written down in formal terms, as rational addiction theories are, and tested more carefully (see Chapter 10 for more examples).

In this chapter we also introduced a biological distinction between wanting vs liking (translated, for *lingua franca*, as decision utility and experienced utility). The fact that these kinds of revealed utility are not always the same was most clearly established by experiments of Berridge and colleagues. Recall that Robinson et al (2005) administering the dopamine agonist L-DOPA (increasing dopamine) restored effective "wanting" in mice bred to be dopamine-deficient. The mice then showed more revealed wanting in a T-maze task (e.g., running faster for reward), but did not show signs of liking the rewards more. Tindell et al (2009) also were able to increase rodents' wanting for salt solutions before there was any direct experienced utility of liking salt.

Scientific progress should help us to establish more about when wanting and liking deviate. This is a major challenge in economics-adjacent social sciences which rest on the foundational principle of consumer sovereignty (or "*de gustibus non est disputandum*" in Becker and Stigler's famous usage, translated as "there's no arguing with tastes"). Consumer sovereignty is often translated and applied, in social behavior and government policy, as not trying to force or limit individual choice freedom, based on the conceit that people don't always want what they will like, and that parents or regulators know better.

In this chapter, we also learned about going beyond the simplest two-stage problems described in chapter 3. Most human choices involve many options and sequences of sub-choices that can be expressed as a decision tree. Something as mundane as going out to eat requires choosing from several restaurants, weighing transportation options, then choosing where to park and sit, then choosing from a large menu.

Simple reinforcement learning is too slow to learn valuations in such trees from experience. A promising alternative is that learning is hierarchical in some way. You might learn categorical structure such as a general value for types of restaurants (Thai, Mexican, steakhouses…), values for walking versus driving, and value for classes of dishes within each restaurant category. Models of hierarchical RL have been imported from computer science into computational neuroscience.

At the same time, modern decision theory in economics is coming to grips with how limited attention and "consideration sets" (subsets chosen from much larger choice sets) can be described.[111] These ideas are approaching hierarchical and categorical structure from a different direction. Furthermore, there are many social science choices which correspond to complex decision trees which have a "scaling problem" (the number of choices explodes combinatorially), such as sequential search for goods or jobs, complex sequences of political choices and vote-swapping, or long-horizon dynamic programming problems such as retirement savings, or household fertility and labor supply decisions. In these cases, the new ideas from decision neuroscience, about how hierarchy and heuristics are used by people, could prove to be insightful.

Beyond addressing these four categories, another novel idea is that similar neuroscientific instantations of habits might also work at higher levels, from household to communities to built organizations. Suppose the (ultimate) goal and value of personal habit MF mode is to shift reliably-rewarding decisions onto an effort-reducing autopilot. Organizations seem to have similar goals, to automate individual-actor decisions to save mental effort and to coordinate activity. Furthermore, habitual automation of routines, language, and cultural practices can improve coordination of activity when "being on the same page" is useful.

Studies along these lines could leverage the kind of work done in social neuroscience, showing that social connections are associated with overlap in functional brain connectivity (e.g. Hyon et al 2021). These measures invite the idea that a successful organization is one in which brain activity is synchronized in some way.

[111] TO DO add some cites here