Conditions for Cognition

Peter Dayan[1]

Draft : August 24, 1994

Abstract

The application of Marr's notions of computational theory to data on classical and operant conditioning suggests computational and neural frameworks into which learning algorithms fit. This allow connections to be made between these algorithms and theoretical notions of what behaviour is appropriate, coming in this case from dynamic programming and Kalman filtering. It also helps to reconcile apparently conflicting results.

1 Introduction

Marr [31] differentiated computational, algorithmic and implementational levels of understanding as one way of abstracting the essence of computations from particular instantiations of them. He defined these as:

Computational	What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it is carried out?
Algorithmic	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?
Implementational	How can the representation and algorithm be realised physically?

Marr's programme for using this division was (at least ostensively) top-down, starting from the appropriate notion of the computation and ending with the neural substrate. However Churchland [6], Churchland and Sejnowski [7] and others have pointed out that one can carve computation more finely than this, and, more importantly, that theories at all levels *co-evolve*. For example, constraints between the levels are provided by the phenomena that are taken as part of the theories at those levels. As the theories change within a level, the constraints must change too. This will generally be crucial for explaining complicated phenomena, as otherwise each level by itself is grossly underconstrained [10].

The long term goal of the work described in this chapter is the application of this notion of the levels to the study of how animals learn to predict events of consequence in their environments and learn to react accordingly. The first two computational levels can be more or less tightly identified with ethological and psychological data. In ethological terms, a major computational-level focus is the appropriateness (optimality, *eg* [30, 56], or adequacy, *eg* [43, 44]) of behaviour of animals under particular sorts of resource or informational constraints. Theoretical psychology, at least in conditioning, has been concerned primarily with algorithmic issues, worrying about learning rules and making somewhat simplistic assumptions about representation (see Gallistel [15] for an eloquent lament on this and interesting alternatives).

Unfortunately, the constraints between levels have largely been ignored. Psychological data on how animals behave and modify their behaviour in well controlled settings should provide information on the 'elements'

[1] Parts of the work described were done in collaboration with Read Montague, Terry Sejnowski and Rich Sutton. Support was from the Canadian NSERC and grants to Geoffrey E Hinton.

out of which their ethological competence must be built – however there are surprisingly few links between psychological and ethological theorising (see [17]). Likewise, apart from the camp of theoretical physiological psychology (see, for example, [16, 18, 19, 20, 37, 35]) and experimental physiological psychology, substantial work, using what seem like ever more rococo behavioural protocols, ignores available constraints from diverse neural implementations.

The tasks for a computational theorist focusing on conditioning are clear – constructing multi-level theories whose algorithms a) have instantiations that are consistent with (and/or predictive of) known neural hardware, and b) can be seen as implementing well-founded computations which it would be felicitous to execute. It is possible to see Grossberg and his colleagues' extensive investigations in conditioning quite like this (see for example [19, 20] and references therein) except that their couching of the computational level is not in terms such as optimising some measure of return, but is instead in terms of a set of more disparate desirable characteristics. Oft cited, for instance, is the importance of avoiding the possibility of confusion between different drives (*eg* victual or sexual hunger).

In this paper, the computational level of the theories comes more from engineering – dynamic programming (DP, [5]) and Kalman filtering (see for example [1]). The next section shows how one body of psychological theory on conditioning [53, 54] was actively interpreted [56, 4] such that it is consistent with DP, a method for optimal control, illustrates the sort of goal-seeking behaviour that this would subserve, and discusses a hypothesis for how it may be implemented in the vertebrate brain using neuromodulatory systems. Section 3 indicates how one can unify this approach with another, apparently conflicting, body of conditioning theory using notions from Kalman filtering [51, 52, 11], a method for the iterative extraction of information from noisy and partial samples. Attentional associabilities of stimuli are treated as learning rates, and are themselves learned. Stimuli are left unattended not because of some limitation to an attentional resource, but rather because learning to them would be inappropriate.

2 Dynamic Programming and Conditioning

Dynamic programming (DP) is an engineering technique which can be used for calculating optimal actions in certain classes of environments. Consider the task of solving a maze, in which there are multiple choice points (places where decisions are required) and some goal which it is desirable to reach as soon as possible. An agent caught in this maze has to perform whole sequences of actions before it reaches the goal. Agents are taken as having *policies*, which are (possibly stochastic) mappings from states to the actions they perform there. In a method for improving behaviour which is called policy iteration, DP learns predictions of how far away different choice points are from the goal based on the current policy. These predictions are used to assign credit or blame to actions directly as the agent performs them, and thus to improve the policy. The prediction part of DP also applies in tasks in which there are no actions, and all an agent can do is to assign values to states in terms of the expectations they can reasonably support of the delivery of rewards or punishments. DP gives us a computational level theory of prediction and optimal control.

Classical and operant conditioning tasks will be treated as being ways of probing the manner in which animals make predictions about the causal texture of their environments [14] and use those predictions to act appropriately. This is not a universally accepted way of looking at these phenomena (a remark which also applies to our shameless conflation of the two types of conditioning), but, at least from a computational perspective, it appears unassailable. Conditioned stimuli (CSs) such as lights or tones are presented just before unconditioned stimuli (USs) such as food or water rewards or small electric shocks, or inform the subject that performing some response (such as pressing a bar) will produce some effect (delivery of food or water, electric current to some part of the brain, or removal of some aversive stimulus). Animals might be shown the relationship between CSs and a US a number of times and then probed for some measure of how much the former predict the latter. Unreasonably, most of the discussion below ignores the difference between appetitive and aversive USs, or rather the difference between appetitive USs and the absence of aversive USs.

One of the first theories of conditioning that made an explicit link between computational and algorithmic levels in the sense above was Uttley's *informon* [55]. Units represent CSs or USs, where what they represent is determined (roughly) by inputs they receive that have fixed weights. Connections between these units have variable weights, and are set according to the negative of the mutual information between input and output, using a learning procedure. Although basing the weight of a connection on the mutual information is computationally sound, and indeed has its reflection in more modern conditioning theory (*eg* Gallistel [15]), the notion that the more correlated an input is with an output, the *lower* the weight is quite counterintuitive and does not seem consistent with known data on plasticity, except in the cerebellum [22], whose output cells are inhibitory.

The theory that has been a major target for experimental investigation in classical conditioning for a long time is due to Rescorla and Wagner [38]. Again, units represent CSs and USs, but now the weights from the former to the latter are set to minimise the errors in predictions on a trial-by-trial basis. For instance if a light is shown in conjunction with food, which itself was not predicted by any of the other stimuli that are present, then there is a prediction error (the difference between the delivered and the expected amount of food) that drives learning and encourages the light to come to predict the food. The behaviour of the Rescorla-Wagner (RW) rule is consistent with some experiments and inconsistent with others (see Dickinson [14] for extensive examples and discussion). Importantly, Sutton and Barto [53] gave the RW rule a sound computational interpretation by pointing out that it is equivalent to the delta rule, a rule for making predictions which is familiar from engineering [58]. This identification permits the large body of theory on the delta rule to be applied directly to the RW rule – illuminating issues such as the convergence of the predictions and the speed of learning in particular circumstances.

Sutton and Barto [53] went on to suggest a variation to RW to take account of certain temporal contingencies – *eg* learning is generally much stronger and more reliable if the conditioned stimuli precede the unconditioned ones rather than vice-versa.[2] It was their subsequent modification of their rule, to what Sutton called a *temporal difference* (TD) rule [49, 50, 54] that, as Watkins showed [56], brought it close to dynamic programming and therefore to a well founded computational level theory.

TD is amply discussed in other references, but briefly it is an embodiment of Samuel's [41] insight that one can bridge arbitrary gaps between stimuli and rewards by enforcing *consistency* in predictions. Samuel was interested in having a system learn to evaluate board positions in the game of draughts (checkers) to estimate which player would win. The trouble is that the information about who actually wins is only available right at the end of the game, and there is a temporal credit assignment problem in evaluating positions early on – it is not clear just from the information at the end at which point things started going wrong (or indeed right). However it is clear that predictions should be consistent with each other – the system should generally evaluate similarly positions that it comes across in adjacent moves. Any inconsistency, *ie* any difference between the predictions from adjacent moves, can drive learning in the same way that conventional prediction errors drive learning in the delta rule.

In terms of conditioning, the equivalent of a game is a whole trial, with the potentially complicated temporal relationships between stimuli and the ultimate reward or punishment. Using TD, the system learns to what extent the different CSs it encounters during the trial predict the ultimate US by enforcing consistency in the predictions each makes. Another way of looking at this is to see that TD rules share with Grossberg's rules [18] the idea that a system's own predictions of reward or punishments (like the evaluations of intermediate positions in a game) are, for the purposes of supporting learning at least, indistinguishable from the actual delivery of reward (like actually winning or losing at the end). This means that a CS that the system has learned predicts reward, can itself support reward learning for other stimuli that can predict it. This is exactly

[2] It turns out [32] that their rule is closely related to one that Konorski published [26] in the year before Hebb made his somewhat more famous suggestion [21].

secondary conditioning, and, given enough trials, allows a system to sort out the predictive relationships between stimuli and rewards even in tasks in which these are temporally very complex.

Another way of looking at TD, which brings out its links to a related method due to Deutsch [13], is to see that it seems that a much better way of solving the temporal credit assignment problem is to work backwards rather than forwards. If the system could only start at the end of a trial, when it knows what return it received, it could much more easily calculate which CSs are suitably predictive. This is of course impossible - TD (and other algorithms which support secondary conditioning) get around this by using only forward learning – in the language of engineering, they learn their evaluations in an asynchronous manner. Just as for delta rules, there is now a body of theory on the conditions and manner of convergence of the TD rules [50, 8, 12, 23], making for a well-founded computational account of this sort of prediction learning.

DP-like theories go on to conflate classical and instrumental conditioning. If the agent is able to perform actions which have some effect on the rewards or punishments that it receives (although the effect may only be apparent after some number of further time steps), then the computational task for it is to choose actions which maximise the rewards or minimise the punishments (this is of course what Samuel [41] was interested in using the board position evaluations for in checkers). In tasks such as the maze, in which the quality of actions is only apparent following whole sequences of them, the agent faces a temporal credit assignment problem to work out which of the actions were appropriate and which not. DP solves this iteratively by reducing this credit assignment problem to the one for prediction, as described briefly above.

Imagine that the agent has some (not necessarily explicit) policy which specifies what actions it chooses at which states. It can use reducing inconsistency as a way of learning to evaluate the proximity of each of the different choice points from the goal under this policy and based on repeated runs through the maze. This works by making the choice points secondary reinforcers whose strengths are functions of these proximities. This information can then be used to change the policy to improve it: if an action takes it from a choice point which is far from the goal to one which is substantially closer, then a new policy which boosts the chances of using that action at the choice point will make each choice point nearer the goal, or at worst, leave it the same distance away. This is an asynchronous form of DP's technique of policy iteration [3].[3]

In summary, TD algorithms that originated from the delta-rule, and were designed to be consistent with data on classical conditioning and to make sense as ways of making predictions, turn out to have a computational level description as implementations of dynamic programming. DP is a method of optimal control, and Watkins [56] explicitly makes the link to optimal foraging. It is a consistent way for information from the environment to control the learning of behaviour. An alternative theoretical framework into which it also fits is Killeen's *Mechanics of the Animate* [25] – the evaluations acquired using TD act like potential functions, the force fields from which can 'drive' behaviour. Altogether, this discussion amounts to half a computational theory, the other half coming from the way in which the algorithms are instantiated in the brain and the changes in this rather abstract description that the resulting constraints will force.

Note that other ways of evaluating policies, *ie* other ways of making the predictions, could equally well be used to change the policies in DP (an example of the independence between Marr's different computational levels). An example is Gallistel's conditioning theory [15], which is built on evidence that animals can calculate the number of presentations of rewards in conjunction with different stimuli, and the number of trials in each case, and divide the one by the other to work out conditional rates of reward. A subsequent calculation uses these rates to assign credit to different conditioned stimuli for predicting the rewards on the basis of a set of computational criteria. No learning rates are needed – the results are immediate on inverting a matrix. Although this has attractive computational power, and could certainly be used as an alternative way of making predictions to change actions, its very power in terms of having no learning rates can be a problem. If the predictions must change, as happens if the experimenter alters the contingencies of rewards in an unsignaled manner or if the agent chooses different actions in the maze, then the lack of a

[3] This description has hidden a large number of details about the exact circumstances under which such asynchronous methods of dynamic programming work. Consult Barto, Sutton & Watkins [4] and Barto, Bradtke & Singh [2] for more details.

learning rate translates into a lack of a forgetting rate. Another principle must be invoked to prevent the animal languishing with incorrect predictions for long periods.

Various issues have been hidden in this discussion. As described, TD will only work if the underlying system generating the stimuli and the consequent rewards can be described as a Markov chain in which each state is signaled by its own unique stimuli and time and action steps are discretised – all three of these are quite restrictive conditions, and there are as yet no generally satisfactory ways around them. The circumstances under which the method of policy improvement described above converges are not presently known, although they are for computationally closely related techniques [56, 57, 23]. A concern that is driven more by behavioural data is the effect of motivation or drive, *eg* the learning that does and does not take place when an animal is hungry or sated, and the behaviours that get favoured.

The most major unrequited concern lies at the algorithmic level, namely representation. Dynamic programming is notoriously damned by a curse of dimensionality which makes it, and the asynchronous variants described above, computationally intractable as the sizes of the problems they tackle grow. Appropriate representations and operations over them can circumvent this – for instance a map of the maze can support very different sorts of planning from the ways that result from asynchronous DP, and there is substantial evidence (see [15] for a survey) that such maps are built and used. The other representational issue is generalisation – the sloth of the learning can be avoided if different states can share information appropriately – which, in terms of the simple learning methods described here, requires that their representations should be partially shared. Although there is some work in this area, *eg* [9], it has not been well explored. Another major omission is the manner in which animals can make predictions about when an event is going to happen as well as that it is going to happen. Although there is plenty of evidence about the nature of short and longer term memories, it is not clear how to integrate this into theories of conditioning in a theoretical parsimonious manner.

The other issues that have been hidden relate more to the remaining half of the complete computational theory, namely the way in which the algorithms are implemented. Although there are those for whom this is irrelevant, it seems likely that the baroque curlicues of the results in the behavioural literature come from constraints whose natural expression lies at the level of implementation. The most glaring issue is the relationship between rewards and punishments, and commensurately between excitation and inhibition. For TD, the former pair are merely positive and negative quantities to be predicted in the conventional way, and which can be directly traded off for each other. Predicting the absence of one of them (as in conditioned inhibition) is just like predicting the other. Various results suggest that this is wrong, or at the very least grossly oversimplified – in fact it seems much more likely that there are different systems that antagonise each other (opponent processes [45] or gated dipoles [18, 20]) for representing positive and negative quantities and that details of the temporal relationship between them may be responsible for some of the more ornate effects.

Following Stein's [47, 48] suggestion that norepinephrine, a neuromodulator which affects network properties and developmental and adult synaptic plasticity, might be involved in a reward system for self-stimulation (in which animals control the delivery of electrical current to particular parts of their brain by performing simple behavioural acts), there has been an enormous volume of experimental work, and somewhat less extensive theoretical work (see, for example, [18]), looking at neuromodulatory systems (dopamine, serotonin, *etc*) in vertebrates and their role in reward and other sorts of learning. There is now something of a consensus that dopamine is involved in reward learning, and particularly the dopaminergic cells in the ventral tegmental area (VTA, one of the two main dopamine-projecting areas in the vertebrate brain, the other being the substantia nigra pars compacta, SNc) and their interconnections with the nucleus accumbens and parts of the limbic system and the prefrontal cortex.

There is not the space here to discuss hypotheses about the involvement of this particular pathway in reward learning, but data showing how the responses of the dopaminergic cells develop whilst a monkey is learning a rewarded, motivated, operant conditioning task [27] are suggestive of their projecting information about prediction errors for TD-like rules [37, 33, 35]. The main target of the SNc is the dorsal striatum, which is thought to play an important role in the occurrent and learned selection of actions, and dopamine is believed to be involved in this. In the operant task, Ljunberg, Apicella & Schultz [27] could see no difference between the firing of the cells in the VTA and the SNc. Although this is weak evidence indeed, it is just what one would expect from TD-based asynchronous policy iteration [3]. For that, the prediction error has two roles; acting as the 'delta' to change weights to improve predictions, and informing the action selection system as to the quality of its action so that it can change its policy accordingly. In fact, in a very simple model of decision making, we have used it in a third fashion – to control ongoing actions using klinokinesis up a learned gradient [34], changing actions if this prediction error signal is negative.

Understanding the constraints from the neural substrate is only at an early stage - many questions only arise from the role the systems are being asked to play in the theory. For instance, there are various suggestions that serotonin (particularly supported by data on its involvement in extinction [46]), acetylcholine, or other neuromodulators or peptides act as opponent processes to dopamine, but the data are still somewhat indeterminate as to just which or how – there are no electrophysiological data equivalent to those on the dopaminergic system [27]. Further, there is only suggestive evidence that dopamine cells are inhibited in their firing if the monkeys are denied reward on a particular trial when they might normally expect it [27, 42]. However, evidence from anatomical, pharmacological and computational studies all appears to be pointing in similar directions as to the involvement of these systems in conditioning - an example of coevolutionary convergence.

3 Kalman Filtering and Conditioning

There is an important area of dispute between theories of conditioning to which we have not yet alluded. The Rescorla-Wagner rule [38] was designed to explain phenomena like blocking [24] in which no learning accrues to a CS such as a tone that could be seen as predicting a US if another CS such as a light is presented at the same time as the tone and if the light has already learned to predict just the amount of US that is provided. The RW rule captures this since there is no error if the predictions made on the basis of the light are correct and therefore there is no learning. However there is an alternative class of theories due to Grossberg (see [18]), Mackintosh [29] and Pearce and Hall [36] which appeals to having the animal pay more or less attention to the conditioned stimuli, making them more or less ready to enter into associations with rewards (*ie* more or less associable) and consequently support more or less learning. The attentional explanation of blocking is rather more subtle – rather than there being no learning because the animal expected all the reward there is, no learning occurs because the animal realises that nothing surprising or unexpectedly interesting happens following the addition of the tone, thus has no reason to pay any attention to the tone, and therefore minimises any learning about the relationship between the reward and the tone.

Different theories make different hypotheses about what controls associabilities. Mackintosh [29] argued that stimuli that are established as good predictors should attract substantial attention (*'good predictors gain power'*); Pearce and Hall [36] suggested that only those that have unusual consequences should support learning (*'anomalous outcomes advance associability'*); Grossberg [18] suggested something akin to Pearce and Hall. Experiments which try to bias attention to test these theories do not seem to decide cleanly between them – different circumstances lead to different outcomes (see [14]). The underlying notion is often that an animal's attention is limited, and therefore its allocation should be managed carefully.

In terms of a computational theory, these different hypotheses are rather like computational suggestions for what should control learning. However they are not embedded in complete theories at a computational level, and therefore want for computational constraints. The obvious place to turn for such a computational theory, at least for prediction, is Kalman filtering [1]. This is a technique from engineering which says how systems faced with multiple samples from a noisy process, each of which can tell only part of the story about the current relationship (the weights) between inputs (the CSs) and outputs (the USs), can optimally iteratively improve their estimates. Some recent work by Sutton [51, 52] on how a system might set learning rates for conditioned stimuli in the delta rule turns out to offer a way to unify the various different conditioning theories [11], at least given a number of somewhat unrealistic simplifying assumptions about the nature of an animal's uncertainty about the relationship between stimuli and reward and the way that it assumes this relationship to change stochastically over time.

Briefly, the system learns learning rates (associabilities) for different stimuli by fitting a linear model of the net prediction errors at any time to the presence of the stimuli.[4] Learning the weights in the Kalman filter is always driven by a delta-like term, which gives the relationship to the RW rule, but depending on the circumstances, either Mackintosh-like or Pearce and Hall-like behaviour can result from the learned learning rates. If the system expects the presentation of a CS and US on each trial to be somewhat uncertain, then it will choose to take a number of trials to learn to make perfect predictions to avoid the possibility of being fooled by noise. During the period of learning, it will see that the presence of the CS is correlated with prediction error, which in turn will boost the learning rate for that stimulus – 'good predictors gain power'. However once predictions are perfect, there will be no more prediction errors, and so learning rates of stimuli that are presented will get small and it will take some prediction error to increase them – 'anomalous outcomes advance associability'. Note that there is no appeal to any limitations on a resource – stimuli support little learning if this learning is expected to be worthless or even positively harmful.

Various details have yet to be worked out [11] – for example should prediction errors of all sorts simply be lumped together in calculating associabilities? Although this would seem to be unfortunate since the system would be in grave danger of spending effort predicting large amounts of unpredictable noise, there is actually evidence that something like this happens in phenomena such as that latent inhibition (the effect that a conditioned stimulus presented a large number of times on its own is subsequently much slower to learn to predict a reward than one never shown) is attenuated if each time the conditioned stimulus is presented it is followed by a different, novel, conditioned stimulus, none of which predicts the reward itself. Mere prediction of novelty seems to suffice to maintain the associability of a stimulus.

Other constraints should again come from studying the neural substrate. Although ample evidence implicates the noradrenergic system in the processing of novel or unexpected events, removal of the dorsal noradrenergic ascending bundle (DNAB), the major source of norepinephrine in the cortex, apparently has little or no effect on latent inhibition [40] or blocking [39], whereas manipulation of the processing of dopamine and serotonin affects at least the former [28]. Just how a single system can be partly responsible for processing both rewards and attention or salience is not yet clear.

4 Discussion

Just as interest in the anatomical, physiological and pharmacological substrates of conditioning is waxing, work on algorithms modelling it is waning. This seems more a product of restricted computational horizons than of anything intrinsic in the field, and this paper has tried to argue that by expanding the set of constraints by appealing to computational notions of prediction and optimal control and also to evidence from the neural substrates, new light can be shed on the phenomena and their role in the cognitive economy.

In particular, the theory of dynamic programming, its asynchronous algorithmic instantiation using temporal difference learning, and its potential dopaminergic substrate has been suggested as a way of understanding how predictions of outcomes are made in temporally complex conditioning tasks and how these predictions can be used to achieve worthwhile ethological goals. Likewise, a rather less well worked-out theory has been

[4] This algorithm is due to Sutton [51, 52] and involves making a diagonal approximation to the matrix of covariances of the uncertainties in the estimates of the weights, and then learning these diagonal values.

postulated of how Kalman filtering and its potential algorithmic and neural instantiations might mediate learned changes in associabilities or learning rates, again in a computationally justified way.

These two theoretical strands need to be wed and also require substantial fleshing out in terms of experiments at various levels. However they should provide the multi-level foundations for the more elaborate and more effective representations and computations implied by copious experiments.

References

[1] Anderson, BDO & Moore, JB (1980). *Optimal Filtering.* Eaglewood, NJ: Prentice-Hall.

[2] Barto, AG, Bradtke, SJ & Singh, SP. *Real-Time Learning and Control using Asynchronous Dynamic Programming.* TR 91-57, Department of Computer Science, University of Amherst, MA. 1992.

[3] Barto, AG, Sutton, RS & Anderson, CW (1983). Neuronlike elements that can solve difficult learning problems. *IEEE Transactions on Systems, Man, and Cybernetics,* **13**, pp 834-846.

[4] Barto, AG, Sutton, RS & Watkins, CJCH (1989). *Learning and Sequential Decision Making.* Technical Report 89-95, Computer and Information Science, University of Massachusetts, Amherst, MA.

[5] Bellman, RE. *Dynamic Programming.* Princeton, NJ: Princeton University Press, 1957.

[6] Churchland, PS (1986). *Neurophilosophy.* Cambridge, MA: MIT Press, Bradford Books.

[7] Churchland, PS & Sejnowski, TJ (1988). Perspectives in cognitive neuroscience. *Science,* **242**, 741-745.

[8] Dayan, P (1992). The convergence of TD(λ) for general λ. *Machine Learning,* **8**, 341-362.

[9] Dayan, P (1993). Improving generalisation for temporal difference learning: The successor representation. *Neural Computation,* **5**, 613-624.

[10] Dayan, P (1994). Computational modelling. *Current Opinion in Neurobiology,* **4**, 212-217.

[11] Dayan, P, Montague, PR, Sejnowski, TJ & Sutton, R. *Predicting Surprise.* In preparation.

[12] Dayan, P & Sejnowski, TJ (1994). TD(λ): Convergence with probability 1. *Machine Learning,* in press.

[13] Deutsch, JA (1960). *The Structural Basis of Behavior.* Chicago, Il: University of Chicago Press.

[14] Dickinson, A (1980). *Contemporary Animal Learning Theory.* Cambridge, England: Cambridge University Press.

[15] Gallistel, CR (1990). *The Organisation of Learning.* Cambridge, MA: MIT Press, Bradford Books.

[16] Gluck, MA & Thompson, RF (1987). Modelling the neural substrates of associative learning and memory: A computational approach. *Psychological Review,* **94**, 176-191.

[17] Gould, JL & Marler, P (1984). In P Marler & HS Terrace, editors, *The Biology of Learning.* Berlin: Springer-Verlag.

[18] Grossberg, S, editor (1987). *The Adaptive brain.* Amsterdam: North-Holland.

[19] Grossberg, S & Levine, DS (1987). Neural dynamics of attentionally modulated Pavlovian conditioning: Blocking, inter-stimulus interval, and secondary reinforcement. *Applied Optics,* **26**, 5015-5030.

[20] Grossberg, S & Schmajuk, NA (1987). Neural dynamics of attentionally modulated Pavlovian conditioning: Conditioned reinforcement, inhibition, and opponent processing. *Psychobiology,* **15**, 195-240.

[21] Hebb, DO. *The Organization of Behavior: A Neuropsychological Theory.* New York, NY: Wiley, 1949.

[22] Ito, M (1989). Long-term depression. *Annual Review of Neuroscience.* **12**, 85-102.

[23] Jaakkola, T, Jordan, MI & Singh, SP (1994). Stochastic convergence of iterative dynamic programming algorithms. Submitted to *Neural Computation.*

[24] Kamin, LJ (1969). Predictability, surprise, attention and conditioning. In BA Campbell & RM Church, editors, *Punishment and Avoidance Behavior,* 276-296. New York, NY: Appleton-Century-Crofts.

[25] Killeen, PR (1992). Mechanics of the animate. *Journal of the Experimental Analysis of Behaviour,* **57**, 429-463.

[26] Konorski, J. *Conditioned Reflexes and Neuron Organisation.* Cambridge, CUP, 1948.

[27] Ljunberg, T, Apicella, P & Schultz, W (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology,* **67**(1), 145-163.

[28] Loskutova, LV, Luk'yanenko, FYa, Il'yuchenok, RYu (1990). Interaction of serotonin- and dopaminergic systems of the brain in mechanisms of latent inhibition in rats. *Neuroscience and Behavioral Physiology,* **20**, 500-505.

[29] Mackintosh, NJ (1975). A theory of attention: Variations of the associability of stimuli with reinforcement. *Psychological Review,* **82**, 276-298.

[30] Mangel, M & Clark, CW (1988). *Dynamic Modeling in Behavioral Ecology.* Princeton, NJ: Princeton University Press.

[31] Marr, D (1982). *Vision.* New York, NY: Freeman.

[32] McLaren, IPL & Dickinson, A (1990). The conditioning connection. *Philosophical Transactions of the Royal Society of London, B,* **329**, 179-186.

[33] Montague, PR, Dayan, P, Nowlan, SJ, Pouget, A & Sejnowski, TJ (1993). Using aperiodic reinforcement for directed self-organization. In CL Giles, SJ Hanson & JD Cowan, editors, *Advances in Neural Information Processing Systems, 5,* 969-977. San Mateo: CA: Morgan Kaufmann.

[34] Montague, PR, Dayan, P & Sejnowski, TJ (1994). Foraging in an uncertain environment using predictive Hebbian learning. In JD Cowan, G Tesauro & J Alspector, editors, *Advances in Neural Information Processing 6.* San Mateo, CA: Morgan Kaufmann.

[35] Montague, PR & Sejnowski, TJ (1994). The predictive brain: Temporal coincidence and temporal order in synaptic learning. *Learning and Memory,* in press.

[36] Pearce, JM & Hall, G (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review,* **87**, 532-552.

[37] Quartz, SR, Dayan, P, Montague, PR & Sejnowski, TJ (1992). Expectation learning in the brain using diffuse ascending projections. *Society for Neuroscience Abstracts,* **18**, 1210.

[38] Rescorla, RA & Wagner, AR (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. In AH Black & WF Prokasy, editors, *Classical Conditioning II: Current Research and Theory,* 64-69. New York, NY: Appleton-Century-Crofts.

[39] Robbins, TW, Everitt, BJ, Cole, BJ, Archer, T & Mohammed, A (1985). Functional hypotheses of the coeruleocortical noradrenergic projection: A review of recent experimentation and theory. *Physiological Psychology,* **13**, 127-150.

[40] Robbins, TW, Everitt, BJ, Fray, PJ, Gaskin, M, Carli, M & de la Riva, C (1982). The roles of the central catecholamines in attention and learning. In MY Spiegelstein & A Levy, editors, *Behavioral Models and the Analysis of Drug Action,* 109-134. Amsertdam: Elsevier.

[41] Samuel, AL (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development,* **3**, pp 211-229.

[42] Schultz, W, Apicella, P & Ljungberg, T (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience,* **13**, 900-13.

[43] Simon, HA (1955). A behavioral model of rational choice. *Quarterly Journal of Economics,* **69**, 99-118.

[44] Simon, HA (1956). Rational choice and the structure of the environment. *Psychological Review,* **63**, 129-138.

[45] Solomon, RL & Corbit, JD (1974). An opponent process theory of motivation, I: Temporal dynamics of affect. *Psychological Review,* **88**, 119-145.

[46] Soubrie, P (1986). Reconciling the role of central serotonin neurons in human and animal behavior. *Behavioral and Brain Sciences,* **9**, 319-335.

[47] Stein, L (1962). Effects and interactions of imipramine, chlorpromazine, reserpine and amphetmaine on self-stimulation: Possible neurophysiological basis of depression. In J Wortis, editor, *Recent Advances in Biological Psychiatry,* 228-308. New York, NY: Plenum Press.

[48] Stein, L (1968). Chemistry of reward and punishment. In DH Efron, editor, *Psychopharmacology. A Review of Progress,* 105-123. Washington, DC: GPO.

[49] Sutton, RS (1984). *Temporal Credit Assignment in Reinforcement Learning.* PhD Thesis, University of Massachusetts, Amherst, MA.

[50] Sutton, RS (1988). Learning to predict by the methods of temporal difference. *Machine Learning,* **3**, pp 9-44.

[51] Sutton, RS (1992a). Adapting bias by gradient descent: An incremental version of delta-bar-delta. In *Proceedings of AAAI-92.*

[52] Sutton, RS (1992b). Gain adaptation beats least squares? In *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems.*

[53] Sutton, RS & Barto, AG (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review,* **88** 2, pp 135-170.

[54] Sutton, RS & Barto, AG (1987). A temporal-difference model of classical conditioning. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society.* Seattle, WA.

[55] Uttley, AM (1976). A two-pathway informon theory of conditioning and adaptive pattern recognition. *Brain Research,* **102**, 23-35.

[56] Watkins, CJCH (1989). *Learning from Delayed Rewards.* PhD Thesis, University of Cambridge, England.

[57] Watkins, CJCH & Dayan, P (1992). Q-learning. *Machine Learning,* **8**, 279-292.

[58] Widrow, B & Stearns, SD (1985). *Adaptive Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall.