Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Neurobiology

How to set the switches on this thing

Peter Dayan

Reinforcement learning (RL) has become a dominant computational paradigm for modeling psychological and neural aspects of affectively charged decision-making tasks. RL is normally construed in terms of the interaction between a subject and its environment, with the former emitting actions, and the latter providing stimuli, and appetitive and aversive reinforcement. However, there is recent emphasis on redrawing the boundary between the two, with the organism constructing its own notion of reward, punishment and state, and with internal actions, such as the gating of working memory, being treated on an equal footing with external manipulation of the environment. We review recent work in this area, focusing on cognitive control.

Address
Gatsby Computational Neuroscience Unit, University College London, 17 Queen Square, London WC1N 3AR, United Kingdom

Corresponding author: Dayan, Peter (dayan@gatsby.ucl.ac.uk)

Current Opinion in Neurobiology 2012, **22**:1068–1074

This review comes from a themed issue on **Decision making**

Edited by **Kenji Doya** and **Michael N Shadlen**

For a complete overview see the Issue and the Editorial

Available online 15th June 2012

0959-4388/$ – see front matter,

http://dx.doi.org/10.1016/j.conb.2012.05.011

Introduction

The theory of Bayesian decision making is formally very straightforward. Its components are a probability distribution over the possible states of the world, a set of available actions, and the value associated with each action under each state. The action should be chosen which maximizes the mean value under the current state of the world. This can be applied in everything from the simple decision problems associated with signal detection theory [1], to spatially and temporally complex tasks associated with structured environments (for instance, partially observable Markov decision processes in which information about past sensory inputs is necessary to disambiguate the current state [2]).

Reinforcement learning (RL; [3]) encompasses a powerful set of computational ideas and methods for organizing and solving such decision-theoretic problems. It has particularly rich links with the psychology of animal conditioning and the neuroscience of appetitive choice. Indeed, as abundant reviews (many in the pages of this journal) have noted, experimental questions in this area are frequently couched (albeit not necessarily resolved) in terms deriving partly from RL.

The bulk of work in neural RL assumes that the state of the world is known transparently. It focuses on the acquisition of knowledge that allows the values of actions in each state to be predicted or computed, along with algorithms that can carry out the computations concerned, leading to value-dependent action choice. Neural RL encompasses various techniques, particularly for domains in which sequences of actions or trajectories must be chosen on the basis of their long-run returns.

One notable division is between model-based (or goal-directed) and model-free (or habitual) methods [4–6]. A model of an environment predicts the transitions between states that will occur when each action is taken, and also the rewards that are likely to result from those transitions and states. Model-based methods build such a model, and then compute long-run values by using it to predict the accumulated future reward associated with possible trajectories [5,4]. By contrast, model-free methods learn to predict the same quantities, that is, the expected sums of such future rewards along trajectories, but without building a model. Instead they take advantage of the fact that predictions of long-run values from states that are encountered successively should be consistent with other. They should only differ according to the reinforcement delivered through the transition. Model-free learning uses any inconsistency as a form of temporally sophisticated prediction error that can correct the predictions.

Model-based and model-free methods have different computational and statistical properties, and are instantiated in at least partially different neural tissue [6–8]. There is also Pavlovian control, in which actions are automatically elicited in the light of reinforcers or predictions of those reinforcers, whether or not those actions are actually contingent, or even beneficial, for acquiring the rewards or avoiding the punishments [9,10]. There are various ways in which Pavlovian influences are exerted, including being embedded in the neural architecture of choice (for instance, differential neuromodulation on direct (go), versus indirect (no-go) pathways through the striatum; [11]), and via other manipulations of such architectural elements [12].

RL is often used to model a rather restricted set of decision problems. However, many others can also usefully be embraced. Consider, for instance, the broad set of

tasks involving the integration of noisy sensory information over time to determine an appropriate choice. A wealth of psychological and neural investigations has centred around one class of policies — namely diffusion-to-bound decision making, associated with sequential probability ratio tests [13,14]. These and their extensions have provided sharp hypotheses for understanding everything from speed-accuracy trade-offs [15,16] to competition in pools of interacting units [17], to the activity of neurons in area LIP during information integration [18]. From the perspective of RL, such sequential probability ratio tests offer a formally beautiful, but brittle, policy for acting in a very restricted class of partially observable Markov decision problem [19]. As soon as some basic conditions are violated, for instance if the signal to noise ratio of the information being integrated is variable or there is a short deadline for producing a response, broader ideas must be used, such as theory derived from RL principles. This observation has been used to capture neurophysiological observations such as urgency signals that have been seen in LIP [20•]. It has even been extended as far as information integration in schizophrenia [21].

Nevertheless, we will argue that even this expanded view of RL is limited, because it retains a neat boundary between organism and environment. We briefly review recent work that offers extensions to all three components of Bayesian decision theory: the definition of value, the nature of the state; and closely related to this, the menu of possible actions. First, one of the claims of the field of intrinsically motivated RL [22•] is that value lies within the subject and not outside. That is, there is no pathway communicating positive or negative values directly from the environment, it is only stimuli of different sorts that are received. Second and third, the internal neural computations that result in state representations and external actions can themselves be described as involving decisions — for example, whether or not to preserve a stimulus in (metabolically expensive) working memory [23,24•,25•,26•], how to set parameters associated with speed-accuracy tradeoffs [27], or even how deep a tree of states and actions to build in order to assess model-based choice [28].

When is a reward a reward?

One of the central ideas in the field of intrinsic motivation [29] is the apparently postmodern notion that rewards and punishments are constructions of the subject rather than products of the environment. One reason for this is to license factors such as curiosity [30] or play as determining appropriate behaviour, even though, on the face of it, they would seem to work against maximizing reward. This reason can in fact be criticized [31] on the grounds that curiosity is actually a perfectly normal facet of control in the face of uncertainty — rewards associated with it are forms of bonus [32,33] that allow agents to trade exploration of potential opportunities afforded by a partially known environment against exploitation of existing knowledge. Certain observing [34] or information seeking [35] actions that provide data on the state *in* the world (e.g. the location of the agent in a maze) rather than the state *of* the world (the overall connectivity of the maze) can also be seen in this light, although it is then less clear why these would persist even when they are clearly deleterious [36]. Importantly, the uncertainty about the world that is associated with curiosity and exploration is always relative to the subject's knowledge. Thus it introduces an internal, inevitably subjective, element to signals associated with reward.

The second reason for the intrinsic nature of reward is that it is best seen as being in the eye of the beholder. The environment might afford different sorts of nutrient or fluid depending on actions; whether these are rewarding is a function of the nature and state of the subject. Indeed, in the fields of psychology and neuroscience, the dependence of reward on the motivational state of the subject has of course long been recognized (see [37] for discussion), including such extreme examples as top-down modulation associated with the exertion of self-control [38]. Indeed that model-based and model-free systems are differentially sensitive to motivational manipulations underpins many of the tests seeking to discriminate which is in control over behaviour [6,39]. Even brain stimulation reward shows motivational-state dependence as a function of electrode placement [40], arguing for a topographic mapping of motivational factors.

Critically, the seemingly clean normativity of the familiar description of decision-making problems sketched at the outset is complicated by this flexibility and at least three other major factors. First, many subjects have a particular difficulty in predicting the values that outcomes will have under motivational states they do not currently occupy [41•]. Second, subjects find it challenging to create unitary values from multidimensional characteristics of the outcomes of choices (even when it is as simple as only having different probabilities and magnitudes), and thus to realize a stable basis for comparison [42]. Third, stability is also challenged by the observation that values adapt rapidly to local statistics [43]. With the value aspect of Bayesian decisions being evidently so labile, the key task ahead lies at the next level up — determining the principles of the mutability itself.

Who controls the controllers?

In decision-theoretic terms, notions such as intrinsic motivation redraw the boundary between organism and environment, placing the values of choices in the former rather than the latter. How about other components of a decision problem, namely the nature of state and the menu of possible actions?

State is indeed itself an intrinsic rather than extrinsic construct. This has a critical impact in at least two rather

opposite respects. First, organisms are bombarded by huge amounts of information, only a tiny fraction of which is relevant to any decision being made. This means that they face a problem of filtering or attention [44,45•], whittling down the data deluge to its meaningful elements. Which elements this should include is, of course, a function of the (possibly incompletely known) task.

Second, despite its bulk, the information in the current sensory input itself is frequently insufficient to decide what to do. Rather, information from the past is necessary to construct a notion of state that is adequately predictive of the values of actions. This is formally very well understood in the context of partially observable Markov decision problems [2], which leads to the solution of storing judiciously selected aspects of the history of past inputs (and possibly past actions) to disambiguate the future. This contingent storage can be readily identified with gated persistent working memory [24•,46].

These two classes of neurocomputational operations for creating the intrinsic state were described in terms of decisions themselves. That is, focusing on one input channel, or storing a stimulus in working memory, are elective choices. These choices have an impact on the efficacy of the organism's actions in its environment, and so the rewards and punishment it receives. It therefore becomes compelling to think [25•] of these, and indeed many other, internal choices associated with cognition itself in just the same terms as external actions. Internal choices may also be associated with a set of additional computational costs [47•], for instance if persistent activity underlying working memory is metabolically expensive.

This overall claim has two facets, one descriptive; the other mechanistic. The weaker, descriptive, notion is that it is helpful to consider the problems of cognition itself in decision-theoretic terms. Given structural and functional assumptions, such as the (actually notional; [48]) capacity of working memory, or the time it takes to gate information into it, it is possible to quantify the costs and benefits of particular neural algorithms. We could thus assess the quality of fit of subjects to their cognitive environments, and, concomitantly, how this environment is fit to the external decision-making problems [49•]. Ideas like this are abroad, for instance in trying to understand task switching (MT Todd, PhD thesis, Princeton University, 2012), where the costs of maintaining a prepared task might tip the balance between employing proactive (i.e. pre-prepared) versus reactive (i.e. unprepared) control [50]. Normative formulations of paradigms such as the stop-signal reaction time task [51] should facilitate the process of uncovering the functional constraints and costs that limit prompt and flexible actions. 'Meta-learning', that is, the setting of parameters in such policies as diffusion-to-bound decision-makers [27,16], or parameters that control exploration [52] can be similarly accommodated. One could also at least superficially encompass classes of policy informed by concepts from supervised, rather than reinforcement, learning [53].

The stronger, mechanistic, claim is that the methods for learning and expressing policies associated with RL (potentially including model-free, model-based and Pavlovian processes) are as applicable to internal as external actions. The seminal suggestions along these lines were originally made by Todd Braver, Jonathan Cohen, Michael Frank, Randy O'Reilly, and their colleagues, particularly focused on the role of working memory in cognitive control [54,55•,24•,56,25•,23,57]. They consider an architecture in which prefrontal cortex can manipulate sensory and motor processing in other parts of the brain on the basis of information associated with the task that is maintained in persistent activity. Roughly, storing information changes the internal state, and so changes the prevailing mapping from input to external actions, and also internal ones. Effects on the future trajectory of external actions lead to rewards and punishments that can criticize and thus improve the original, internal, storage action.

One of the test-beds for this notion has been the so-called 12AX task [24•]. Subjects see a sequence of letters or an occasional number '1' or '2'. They have to respond in a special way at the end of the subsequence 'AX' if the most recent number they saw was a '1'; or at the end of 'BY' if the most recent number they saw was a '2'. In terms of gated working memory, this can be solved using two levels of storage — the most recent number ('1' or '2'), and depending on this, the last letter if it was 'A' or 'B'. This then enables the final action to be chosen appropriately given a subsequent 'X' or 'Y'. This is a simple example of a partially observable problem — it is impossible to decide what to do given only the currently observed letter or number. The original investigations of this task involved a hard-wired policy with internal and external actions that ran along the lines of gating [24•]; more recently, it has been shown that modified [25•] and standard [26•] forms of model-free RL can be used to acquire such a policy. Gated working memory has been extended to hierarchical RL [58–61], with segregated cortico-striatal circuits interacting with each other in a layered manner. This provides a link to work on structural hierarchies in the realization of cognitive control by dorsal areas of prefrontal cortex [62,63•,64–67]. In turn, these are tied with psychological ideas about hierarchical control schemas [68] and general theories of computational [69] and neural [70,71•] hierarchical RL. These latter methods actually realize a third form of intrinsic reward, associated with the attainment of subgoals, that is, the way that more abstract, higher-level, actions criticize more concrete, lower-level ones [72•].

Much of this work involves model-free forms of RL, which certainly seems more obviously to avoid any thorny regresses associated with explaining complex methods for choosing external actions with equally complex methods for choosing internal actions. However, one pressing direction for future research is to understand in a hierarchical architecture how model-free policies might realize model-based calculations of values, and even of the uncertainties about those values that putatively determine arbitration between model-based and model-free choice [5]. This would then provide a substrate for model-based selection of both internal and external actions. It could also inform the workings of other associates of model-based control [73], such as instructed control [74,75], or indeed top-down control in the context of cognitive architectures such as ACT-R [76,77]. In a more elaborated hierarchy, it might even be that model-free control at one level depends on model-based control at a lower level.

Tasks exploring the relationship between model-free and model-based control [78–80] and the perversion or palliation of either by putative Pavlovian influences [81] are under active development. One interesting prospect is that in these interactions will be found the basis of new ways of looking at neurological and psychiatric dysfunction [82,83].

A rather separate direction for the instantiation of model-based control comes from its reduction to inference in a special form of belief net [84,85]. The idea is that the very same probabilistic computational operations that hierarchically arranged sensory processing cortical areas are believed to do to model and represent complex input, might also be performed by hierarchically arranged prefrontal and premotor cortical areas to model and realize complex behavioural plans. This certainly offers a much more straightforward link to some aspects of hierarchical control [70], and thus to a deeper understanding of the functional structure of cognitive control [71•].

Conclusions

Abbott [86] titled a chapter in a book on problems in systems neuroscience 'Where are the Switches on This Thing'. His examples were drawn from neural systems other than those we have considered; however, he was really asking about how what we have considered as cognitive actions such as selection and gating are realized, by neuromodulation (dismissed there to rather short order, although favoured by some of the models here; [55•]), inhibition or gain control.

In these terms, we have considered instead how such switches should be set. We considered new thinking on both senses of 'how': what the switch-setting should achieve; and the process by which the switches might come to be set correctly. These involved redrawing the interface between inside the agent and outside in the environment, placing reward inside, even though it is normally outside, but treating the switches as if they are outside, when they are normally inside.

Acknowledgements

I am very grateful to Matt Botvinick, Zeb Kurth-Nelson, Randy O'Reilly and the editors for comments on an earlier version. Funding was from the Gatsby Charitable Foundation.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Green DM, Swets JA: *Signal Detection Theory and Psychophysics*, vol 1974, New York: Wiley; 1966.
2. Kaelbling LP, Littman ML, Cassandra AR: **Planning and acting in partially observable stochastic domains**. *Artif Intell* 1998, **101**:99-134.
3. Sutton RS, Barto AG: *Reinforcement Learning: An Introduction*. MIT Press; 1998.
4. Doya K: **What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?** *Neural Netw* 1999, **12**:961-974.
5. Daw ND, Niv Y, Dayan P: **Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control**. *Nat Neurosci* 2005, **8**:1704-1711 http://dx.doi.org/10.1038/nn1560.
6. Dickinson AD, Balleine BW: **The role of learning in the operation of motivational systems**. In *Stevens' Handbook of Experimental Psychology*, vol 3. Edited by Randy G. New York: Wiley; 2002:497-534.
7. Killcross S, Coutureau E: **Coordination of actions and habits in the medial prefrontal cortex of rats**. *Cereb Cortex* 2003, **13**:400-408.
8. Balleine BW: **Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits**. *Physiol Behav* 2005, **86**:717-730 http://dx.doi.org/10.1016/j.physbeh.2005.08.061.
9. Breland K, Breland M: **The misbehavior of organisms**. *Am Psychol* 1961, **16**:681-684.
10. Dayan P, Niv Y, Seymour B, Daw ND: **The misbehavior of value and the discipline of the will**. *Neural Netw* 2006, **19**:1153-1160 http://dx.doi.org/10.1016/j.neunet.2006.03.002.
11. Frank MJ, Claus ED: **Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal**. *Psychol Rev* 2006, **113**:300-326 http://dx.doi.org/10.1037/0033-295X.113.2.300.
12. Murschall A, Hauber W: **Inactivation of the ventral tegmental area abolished the general excitatory influence of Pavlovian cues on instrumental performance**. *Learn Mem* 2006, **13**:123-126 http://dx.doi.org/10.1101/lm.127106.
13. Wald A: *Sequential Analysis*. New York: Wiley; 1947.
14. Gold JI, Shadlen MN: **Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward**. *Neuron* 2002, **36**:299-308.
15. Ratcliff R: **Theoretical interpretations of the speed and accuracy of positive and negative responses**. *Psychol Rev* 1985, **92**:212.
16. Bogacz R, Wagenmakers E-J, Forstmann BU, Nieuwenhuis S: **The neural basis of the speed-accuracy tradeoff**. *Trends Neurosci* 2010, **33**:10-16 http://dx.doi.org/10.1016/j.tins.2009.09.002.
17. Usher M, McClelland JL: **The time course of perceptual choice: the leaky, competing accumulator model**. *Psychol Rev* 2001, **108**:550-592.

18. Gold JI, Shadlen MN: **The neural basis of decision making**. *Annu Rev Neurosci* 2007, **30**:535-574 http://dx.doi.org/10.1146/annurev.neuro.29.051605.113038.

19. Dayan P, Daw ND: **Decision theory, reinforcement learning, and the brain**. *Cogn Affect Behav Neurosci* 2008, **8**:429-453 http://dx.doi.org/10.3758/CABN.8.4.429.

20. • Latham PE, Roudi Y, Ahmadi M, Pouget A: **Deciding when to decide**. *Society for Neuroscience*. 2007:. 740.10/YY10.
The sequential probability ratio test (SPRT; [13]) amounts to an optimal policy for a very particular form of speed-accuracy trade-off. This has had a most important impact on the psychology and neuroscience of uncertainty-sensitive decision-making [15,14], but is brittle. This study considers the case that the information bearing on a decision is of unpredictable quality; it is then the more general framework of partially observable Markov decision problems that applies. Other work takes on different ways in which the SPRT is insufficient [51].

21. Moutoussis M, Bentall RP, El-Deredy W, Dayan P: **Bayesian modelling of jumping-to-conclusions bias in delusional patients**. *Cogn Neuropsychiatry* 2011, **16**:422-447 http://dx.doi.org/10.1080/13546805.2010.548678.

22. • Singh S, Lewis RL, Barto AG: **Where do rewards come from**. In *Proceedings of the Annual Conference of the Cognitive Science Society*. 2009:2601-2606.
This paper presents a limpid account of the rationale for treating reward as being intrinsic rather than extrinsic, with the former not following the most obvious form of the latter — that is, fitness.

23. O'Reilly RC, Noelle DC, Braver TS, Cohen JD: **Prefrontal cortex and dynamic categorization tasks: representational organization and neuromodulatory control**. *Cereb Cortex* 2002, **12**:246-257.

24. • Frank MJ, Loughry B, O'Reilly RC: **Interactions between frontal cortex and basal ganglia in working memory: a computational model**. *Cogn Affect Behav Neurosci* 2001, **1**:137-160.
The explicit gating of working memory, an idea borrowed from [46] was a critical step in moving the boundary of reinforcement-learning control inside the agent. This built on the notions of gating based on phasic release of dopamine that was considered in [55], but, taking advantage of the interaction between cortex and striatum, offers a richer and more flexible mechanism that has generalized widely.

25. • O'Reilly RC, Frank MJ: **Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia**. *Neural Comput* 2006, **18**:283-328 http://dx.doi.org/10.1162/089976606775093909.
Having shown that the so-called prefrontal cortex/basal ganglia working memory model could realize sophisticated forms of cognitive control (i.e. appropriate switches are available to perform the relevant tasks), an important question is whether reinforcement learning mechanisms could be used to determine these settings for tasks such as 12AX. This paper suggests a modification of standard reinforcement learning that is able to learn appropriate. Ref [26•] shows that standard methods can do too.

26. • Todd MT, Niv Y, Cohen JD: **Learning to use working memory in partially observable environments through dopaminergic reinforcement**. *Neural Inform Process Syst* 2009, **21**:1689-1696.
The 12AX task imposes some rather particular requirements on working memory. Ref [25•] suggested some extensions and modifications to standard temporal difference learning [3] so that reinforcement learning methods could be used to acquire an appropriate cognitive policy. This paper shows that more standard methods can work too.

27. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD: **The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks**. *Psychol Rev* 2006, **113**:700-765 http://dx.doi.org/10.1037/0033-295X.113.4.700.

28. Baum EB: *What is Thought?* MIT Press; 2003.

29. Ryan RM, Deci EL: **Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being**. *Am Psychol* 2000, **55**:68.

30. Schmidhuber J: **Curious model-building control systems**. *IJCNN*. IEEE Press; 1991:. 1458–1463.

31. Dayan P: **Exploration from generalization mediated by multiple controllers**. In *Intrinsically Motivated Learning in Natural and Artificial Systems.* Edited by Baldassarre G, Mirolli M. Springer-Verlag; 2012.

32. Sutton R: **Dyna, an integrated architecture for learning, planning and reacting**. *Sigart Bull.* 1991, **2**:160-163 ftp://ftp.cs.umass.eduqSqpubqSqanwqSqpubqSqsuttonqSqsutton-91b.ps.gz.

33. Gittins JC: *Multi-Armed Bandit Allocation Indices*, vol 3. John Wiley & Sons Inc.; 1989, 0471920592.

34. Dinsmoor JA: **Observing and conditioned reinforcement**. *Behav Brain Sci* 1983, **6**:693-728.

35. Bromberg-Martin ES, Hikosaka O: **Midbrain dopamine neurons signal preference for advance information about upcoming rewards**. *Neuron* 2009, **63**:119-126 http://dx.doi.org/10.1016/j.neuron.2009.06.009.

36. Spetch ML, Mondloch MV, Belke TW, Dunn R: **Determinants of pigeons' choice between certain and probabilistic outcomes**. *Anim Learn Behav* 1994, **22**:239-251.

37. Niv Y, Joel D, Dayan P: **A normative perspective on motivation**. *Trends Cogn Sci* 2006, **10**:375-381 http://dx.doi.org/10.1016/j.tics.2006.06.010.

38. Hare TA, Camerer CF, Rangel A: **Self-control in decision-making involves modulation of the vmpfc valuation system**. *Science* 2009, **324**:646-648 http://dx.doi.org/10.1126/science.1168450.

39. Holland PC: **Relations between Pavlovian-instrumental transfer and reinforcer devaluation**. *J Exp Psychol Anim Behav Process* 2004, **30**:104-117 http://dx.doi.org/10.1037/0097-7403.30.2.104.

40. Gallistel CR, Beagley G: **Specificity of brain stimulation reward in the rat**. *J Comp Physiol Psychol* 1971, **76**:199-205.

41. • Loewenstein GF, O'Donoghue T: **Animal spirits: affective and deliberative processes in economic behavior**. Technical report, Social Science Research Network Working Papers Series; 2004
One key problem with intrinsic rewards is that they are state dependent, and yet we are apparently poor at assessing what the value will be in states that we do not currently occupy. This is a failure even of model-based reasoning, with harsh consequences.

42. Tversky A: **Intransitivity of preferences**. *Psychol Rev* 1969, **76**:31-48.

43. Chater N, Vlaev I: **The instability of value**. In *Decision Making, Affect, and Learning: Attention and Performance XXIII.* Edited by Delgado M, Phelps EA, Robbins TW. Oxford Univ Press; 2010:81.

44. Zhaoping L: **Theoretical understanding of the early visual processes by data compression and data selection**. *Network* 2006, **17**:301-334.

45. • Gottlieb J, Balan P: **Attention as a decision in information space**. *Trends Cogn Sci* 2010, **14**:240-248 http://dx.doi.org/10.1016/j.tics.2010.03.001.
In our terms, this paper considers the allocation of attention as the setting of a internal switch which has key consequences for sensory information processing, and thus intrinsic state and ultimately reward. Attentional allocation must attract internal costs [47•], but also opportunity costs in terms of other information that is foregone. The potential value of information is also considered in observing tasks [34,35].

46. Hochreiter S, Schmidhuber J: **Long short-term memory**. *Neural Comput* 1997, **9**:1735-1780.

47. • McGuire JT, Botvinick MM: **Prefrontal cortex, cognitive control, and the registration of decision costs**. *Proc Natl Acad Sci U S A* 2010, **107**:7922-7926 http://dx.doi.org/10.1073/pnas.0910662107.
The operation of cognitive control, that is, certain settings of certain switches, is apparently costly — so subjects might reasonably choose internal actions that will minimize the inconvenience.

48. Bays PM, Husain M: **Dynamic shifts of limited working memory resources in human vision**. *Science* 2008, **321**:851-854 http://dx.doi.org/10.1126/science.1158023.

49. • Howes A, Lewis RL, Vera A: **Rational adaptation under task and processing constraints: implications for testing theories of cognition and action**. *Psychol Rev* 2009, **116**:717-751 http://dx.doi.org/10.1037/a0017187.

The notion of bounded rationality has long played an important role in thinking about cognitive systems. This paper reviews a modern line of work into the nature and consequences of factors that limit untrammeled optimality in cognitive control systems. It has close links with theoretical (MT Todd, PhD thesis, Princeton University, 2012) [71•] and empirical [47•] approaches consonant with reinforcement learning views.

50. Braver TS: **The variable nature of cognitive control: a dual mechanisms framework**. *Trends Cogn Sci* 2012, **16**:106-113 http://dx.doi.org/10.1016/j.tics.2011.12.010.

51. Shenoy P, Yu AJ: **Rational decision-making in inhibitory control**. *Front Hum Neurosci* 2011, **5**:48 http://dx.doi.org/10.3389/fnhum.2011.00048.

52. Aston-Jones G, Cohen JD: **An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance**. *Annu Rev Neurosci* 2005, **28**:403-450.

53. Fusi S, Asaad WF, Miller EK, Wang X-J: **A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales**. *Neuron* 2007, **54**:319-333 http://dx.doi.org/10.1016/j.neuron.2007.03.017.

54. Rougier NP, Noelle DC, Braver TS, Cohen JD, O'Reilly RC: **Prefrontal cortex and flexible cognitive control: rules without symbols**. *Proc Natl Acad Sci U S A* 2005, **102**:7338-7343 http://dx.doi.org/10.1073/pnas.0502455102.

55. • O'Reilly RC, Braver TS, Cohen JD: **A biologically based computational model of working memory**. In *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control.* Edited by Mikaye A, Shah P. New York, NY: CUP; 1999:375-411.

A landmark paper that brought together neural reinforcement learning, notably phasic dopamine signals associated with predictors of reward, and gated working memory. This work then developed via explicit striatal gating [24•,25•] to modern accounts based on hierarchical reinforcement learning [70,71•].

56. Hazy TE, Frank MJ, O'Reilly RC: **Banishing the homunculus: making working memory work**. *Neuroscience* 2006, **139**:105-118 http://dx.doi.org/10.1016/j.neuroscience.2005.04.067.

57. Miller EK, Cohen JD: **An integrative theory of prefrontal cortex function**. *Ann Rev Neurosci* 2001, **24**:167-202 http://dx.doi.org/10.1146/annurev.neuro.24.1.167.

58. Reynolds JR, O'Reilly RC: **Developing PFC representations using reinforcement learning**. *Cognition* 2009, **113**:281-292 http://dx.doi.org/10.1016/j.cognition.2009.05.015.

59. Botvinick MM: **Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy**. *Philos Trans R Soc Lond B: Biol Sci* 2007, **362**:1615-1626 http://dx.doi.org/10.1098/rstb.2007.2056.

60. Botvinick MM: **Hierarchical models of behavior and prefrontal function**. *Trends Cogn Sci* 2008, **12**:201-208 http://dx.doi.org/10.1016/j.tics.2008.02.009.

61. Frank MJ, Badre D: **Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis**. *Cereb Cortex* 2011, **22**:509-526 http://dx.doi.org/10.1093/cercor/bhr114.

62. Koechlin E, Ody C, Kouneiher F: **The architecture of cognitive control in the human prefrontal cortex**. *Science* 2003, **302**:1181-1185 http://dx.doi.org/10.1126/science.1088545.

63. • Koechlin E, Summerfield C: **An information theoretical approach to prefrontal executive function**. *Trends Cogn Sci* 2007, **11**:229-235 http://dx.doi.org/10.1016/j.tics.2007.04.005.

This paper offers one of the first computationally inspired mappings of the putative hierarchy of control tasks onto the putative hierarchy of cortical areas in (lateral) prefrontal cortex concerned with control. Other notable contributions include [64,61,66,71•,62]. Exactly what the cortical hierarchy is, however, subject to some debate [67], and there is a dearth of work on the environmental statistics of tasks that would underpin a normative view of a hierarchy [71•,73].

64. Badre D, D'Esposito M: **Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex**. *J Cogn Neurosci* 2007, **19**:2082-2099 http://dx.doi.org/10.1162/jocn.2007.19.12.2082.

65. Badre D, Kayser AS, D'Esposito M: **Frontal cortex and the discovery of abstract action rules**. *Neuron* 2010, **66**:315-326 http://dx.doi.org/10.1016/j.neuron.2010.03.025.

66. Badre D, Frank MJ: **Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: evidence from fmri**. *Cereb Cortex* 2011, **22**:527-536 http://dx.doi.org/10.1093/cercor/bhr117.

67. Reynolds JR, O'Reilly RC, Cohen JD, Braver TS: **The function and organization of lateral prefrontal cortex: a test of competing hypotheses**. *PLoS One* 2012, **7**:e30284 http://dx.doi.org/10.1371/journal.pone.0030284.

68. Cooper RP, Shallice T: **Hierarchical schemas and goals in the control of sequential behavior**. *Psychol Rev* 2006, **113**:887-916 http://dx.doi.org/10.1037/0033-295X.113.4.887 [discussion 917–931].

69. Sutton RS, Precup D, Singh S: **Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning**. *Artif Intell* 1999, **112**:181-211.

70. Botvinick MM, Niv Y, Barto AC: **Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective**. *Cognition* 2009, **113**:262-280 http://dx.doi.org/10.1016/j.cognition.2008.08.011.

71. • Botvinick MM, Cohen JD: **The computational and neural basis of cognitive control: charted territory and new frontiers**. *Cogn Sci* 2012.

Cognitive control was perhaps a surprisingly early focus for work inspired by the connectionist revolution in cognitive science; but nevertheless led to some of the most influential models in its domain. Furthermore, ideas derived from reinforcement learning played a key role in eliminating the homunuculus [57] that lurks around control concepts. This paper offers an historical perspective, but also points towards the impressive new developments around hierarchical reinforcement learning [70] that, in the words of Botvinick himself (personal communication) are aimed at ''assembling a good set of switches in the first place''.

72. • Ribas-Fernandes JJF, Solway A, Diuk C, McGuire JT, Barto AG, Niv Y, Botvinick MM: **A neural signature of hierarchical reinforcement learning**. *Neuron* 2011, **71**:370-379 http://dx.doi.org/10.1016/j.neuron.2011.05.042.

One of the more influential ideas in hierarchical reinforcement learning concerns options [69], which are sophisticated forms of subgoals. Ref [70] considered how options might provide a useful framework for understanding aspects of the structure and function of the prefrontal cortex and internal actions. Here, one key characteristic of options, namely the intrinsic reward prediction error signal consequent on completing a subgoal, was investigated using fMRI.

73. Dayan P: **Goal-directed control and its antipodes**. *Neural Netw* 2009, **22**:213-219 http://dx.doi.org/10.1016/j.neunet.2009.03.004.

74. Doll BB, Jacobs WJ, Sanfey AG, Frank MJ: **Instructional control of reinforcement learning: a behavioral and neurocomputational investigation**. *Brain Res* 2009, **1299**:74-94 http://dx.doi.org/10.1016/j.brainres.2009.07.007.

75. Cole MW, Etzel JA, Zacks JM, Schneider W, Braver TS: **Rapid transfer of abstract rules to novel contexts in human lateral prefrontal cortex**. *Front Hum Neurosci* 2011, **5**:142 http://dx.doi.org/10.3389/fnhum.2011.00142.

76. Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, Qin Y: **An integrated theory of the mind**. *Psychol Rev* 2004, **111**:1036-1060 http://dx.doi.org/10.1037/0033-295X.111.4.1036.

77. Jilk DJ, Lebiere C, O'Reilly RC, Anderson JR: **SAL: an explicitly pluralistic cognitive architecture**. *J Exp Theor Artif Intell* 2008, **20**:197-218.

78. Gläscher J, Daw N, Dayan P, O'Doherty JP: **States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning**. *Neuron* 2010, **66**:585-595 http://dx.doi.org/10.1016/j.neuron.2010.04.016.

79. Daw ND, Gershman SJ, Seymour BS, Dayan P, Dolan RJ: **Model-based influences on humans' choices and striatal prediction errors**. *Neuron* 2011, **69**:1204-1215 http://dx.doi.org/10.1016/j.neuron.2011.02.027.

80. Simon DA, Daw ND: **Neural correlates of forward planning in a spatial decision task in humans**. *J Neurosci* 2011, **31**:5526-5539 http://dx.doi.org/10.1523/JNEUROSCI.4647-10.2011.

81. Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser J: **Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees**. *Public Libr Sci Comput Biol* 2012, **8**:e1002410.

82. Maia TV, Frank MJ: **From reinforcement learning models to psychiatric and neurological disorders**. *Nat Neurosci* 2011, **14**:154-162 http://dx.doi.org/10.1038/nn.2723.

83. Read Montague P, Dolan RJ, Friston KJ, Dayan P: **Computational psychiatry**. *Trends Cogn Sci* 2012, **16**:72-80 http://dx.doi.org/10.1016/j.tics.2011.11.018.

84. Toussaint M, Storkey A: **Probabilistic inference for solving discrete and continuous state markov decision processes**. In *Proceedings of the 23rd International Conference on Machine Learning*. ACM; 2006:945-952.

85. Solway A, Botvinick MM: **Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates**. *Psychol Rev* 2012, **119**:120-154 http://dx.doi.org/10.1037/a0026435.

86. Abbott LF: **Where are the switches on this thing**. In *23 Problems in Systems Neuroscience.* Edited by van Hemmen JL, Sejnowski TJ. Oxford Univ Press; 2006:423-431.