



## 2009 Special Issue

## Goal-directed control and its antipodes

Peter Dayan\*

UCL, 17 Queen Square, London WC1N 3AR, UK

## ARTICLE INFO

## Article history:

Received 18 December 2008  
 Received in revised form  
 26 February 2009  
 Accepted 13 March 2009

## Keywords:

Goal-directed control  
 Habitual control  
 Model-based reinforcement learning  
 Model-free reinforcement learning  
 Declarative control  
 Procedural control  
 Prefrontal cortex  
 Basal ganglia

## ABSTRACT

In instrumental conditioning, there is a rather precise definition of goal-directed control, and therefore an acute boundary between it and the somewhat more amorphous category comprising its opposites. Here, we review this division in terms of the various distinctions that accompany it in the fields of reinforcement learning and cognitive architectures, considering issues such as declarative and procedural control, the effect of prior distributions over environments, the neural substrates involved, and the differing views about the relative rationality of the various forms of control. Our overall aim is to reconnect some presently far-flung relations.

© 2009 Elsevier Ltd. All rights reserved.

## 1. Introduction

Goals and subgoals play rich qualitative roles in the way that we conceive, describe and indeed model behavior. They provide a convenient structure to decompose tricky problems into elemental pieces, and have been the object of much algorithmic attention. However, they are far from the only way to couch or create procedures that solve such problems, and it can be hard to show from just observed behavior whether or not they have a causal rather than a merely descriptive role.

In this paper, we start from a very sharp operationalization of the distinction between goal-directed control and non-goal-directed control coming from studies in animal conditioning (Dickinson, 1985; Dickinson & Balleine, 2002) which has been rendered computationally in terms of the difference between model-based and model-free reinforcement learning (RL; Sutton & Barto, 1998) control policies (Daw, Niv, & Dayan, 2005). Over the course of learning, behavior migrates from being goal-directed to being non-goal-directed. Such an evolution from a more deliberate to more automatic control has many parallels across fields of psychology and artificial intelligence (e.g., Anderson (1982), Anderson and Lebiere (1998), Anderson et al. (2004), Crossman (1959), Fitts (1964), Logan (1988), Newell and Rosenbloom (1981) and Newell (1990)), and our aim is to elucidate some of the links.

In the operationalization from animal conditioning, a goal-directed action is defined as one that is performed because: (a) the subject has appropriate reason to believe that it will achieve a particular goal, such as an outcome; and (b) the subject has a reason to seek that outcome in the first place. The propensity of the subject to select a goal-directed action must therefore be affected by experimental manipulation of either of these two conditions – notably altering the contingency between action and outcome (e.g., by presenting the outcome in the absence of the action), and reducing or enhancing the attractiveness of the outcome (e.g., by poisoning it). Actions that are not affected by these manipulations are, by definition, not goal-directed. In the conditioning literature, they are typically called *habits*, although that is not to say that they are in any way unitary. There is substantial evidence that given only limited experience of a new environment, choices are affected by the manipulations, but over the course of learning, at least some become immune, and therefore transfer from being goal-directed actions to habits. There is also evidence that these two forms of learning progress in parallel rather than serially, since the degree of affectedness after small and large amounts of training can be manipulated by reversible lesions of specific neural areas that have (therefore) been associated with the two systems (Killcross & Coutureau, 2003).

Daw et al. (2005) noted that this description of goal-directed behavior is consonant with a form of *model-based* optimal control. In this, subjects are assumed to have (or to build from experience) a model of the contingencies in the world, including the outcomes associated with each action at each state of the world, and the utility of each outcome. This model defines a tree of states, actions,

\* Tel.: +44 (0) 20 7679 1175.

E-mail address: [dayan@gatsby.ucl.ac.uk](mailto:dayan@gatsby.ucl.ac.uk).

and outcomes, with each level of the tree being associated with a further action in the future. According to model-based control, actions are chosen by searching this model either forwards to the leaves of the resulting tree (working out the ultimate utility consequent on each action) or backwards from the leaves (in this simple case, working out which action leads to the highest utility outcome). Varieties of both methods are standard in artificial intelligence and computer science (Russell & Norvig, 1995).

Unfortunately, in moderate-sized problems, particularly those such as mazes in which the subjects have to take many actions (turning in many directions) before getting an outcome (such as the exit), searching the tree of possibilities to work out which action is best at a location (i.e. state) imposes computationally ruinous demands on working memory and calculation. Intuition for an alternative method in the case of the maze comes from considering one of two possible functions. One, called an optimal state-value function, reports how far each location is from the exit shortest path. The other, called an optimal state-action-value (or  $Q$  function Watkins, 1989), reports how far each location is from the exit for each possible choice of first action at that location, assuming that subsequent choices are optimal. Given either function, most of the complexities of search can be avoided because it is straightforward simply to move from one location to the neighboring location that is closest to the exit (using the state-value function) or to choose the action associated with the shortest distance to the exit (using the  $Q$  function). Indeed, these functions are exactly what the tree search methods produce.

The field of reinforcement learning has focused on finding methods that can acquire such functions from experience in what is known as a *model-free* manner, without requiring the model to be constructed or searched (Sutton & Barto, 1998). Daw et al. (2005) called these model-free methods *cached*, since they use storage or caching of values based on experience as an alternative to online search or calculation. There is a variety of such methods, from the simplest, which acquire nothing more than the identity of the most successful action at a state (almost like instances of successful choices, Logan, 1988; Williams, 1992), to more complex ones based on value estimates (Barto, Sutton, & Watkins, 1990).

One property of model-free methods is that the stored values associated with one state in the environment are not immediately sensitive to changes in contingencies or outcome utilities at other states. Rather, new stored values have to be acquired from new experience as this filters through chains of transitions between these states. Thus, for instance, subjects might have observed that an outcome has been poisoned, and therefore has low utility, but nevertheless have stored values at distant states that are inconsistent with this fact. Thus, their propensity to choose an action at one of those states that would lead to the outcome would be undiminished. This is exactly the mark of a non-goal-directed action, i.e. a habit.

Daw et al. (2005) made the Bayesian suggestion that the choice between model-based and model-free methods should depend on their relative uncertainties. Model-based methods learn more efficiently than model-free methods, because they use calculation to propagate information around the tree of states, action and outcomes, rather than experience in the world. However, these calculations present tremendous challenges to the neural substrate, implying inaccuracies in the values that are produced. Model-free methods, though inefficient at learning in new environments, present only minor challenges in use. Therefore model-based methods are often less uncertain than model-free methods at the start of learning, but more uncertain after model-free methods have had sufficient experience in a domain. This is one account of the experimental observation of the transition from goal-directed to habitual control.

The precision of the distinction between the two forms of control in conditioning makes it an ideal foil for the discussion

in the paper of issues and generalizations of the notion of goal directed control. We briefly discuss (i) declarative versus procedural control, and (ii) interdependence and independence; (iii) interpreted versus compiled; and (iv) prior-bound versus data-bound characterizations of policies; (v) instructed versus learned control; (vi) prefrontal versus basal ganglia substrates; and finally (vii) reflexive versus deliberative control ('system 1' and 'system 2' in the sense of Kahneman and Frederick (2002) and Stanovich and West (2002)). As we will see, these issues are intertwined. Our overall aim is to put together, and thereby provide a broader context for, related strands of research.

## 2. Antipodes

### 2.1. Declarative versus procedural control

One key distinction between model-based and model-free RL control is that the former is *declarative* in the senses both of multiple memory systems (e.g., Squire, 2004) and programming languages such as PROLOG. That is, the model provides a set of (semantic) facts about the structure of the environment and the subject in the form of a forward or generative model (Dayan, Hinton, Neal, & Zemel, 1995; Kawato & Wolpert, 1998; Wolpert, Ghahramani, & Jordan, 1995). These facts *imply* the optimal choice of action, but they do not by themselves provide an immediate mechanism for *calculating* this optimal choice. Indeed, there are different methods (such as dynamic programming's policy and value iteration algorithms, Bertsekas, 2007; Puterman, 2005) that can perform the search.

By contrast, model-free control is typically *procedural* in the sense of memory systems (Squire, 2004) or *imperative* in the sense of programming languages. That is, it specifies directly the choice of action at each state or location as an imperative command, and is the inverse or recognition model to its generative counterpart (Dayan et al., 1995; Wolpert & Kawato, 1998).

Paired generative and recognition models have substantial currency in unsupervised learning models of sensory plasticity and inference (Dayan et al., 1995; Kawato, Hayakawa, & Inui, 1993; Rao, Olshausen, & Lewicki, 2002), and indeed the analogy between these aspects of sensation and action has been made before (Todorov, 2007). The difficulty of inverting the generative model in an on-line manner, which is exactly analogous to the difficulty of searching in the tree of actions and outcomes, led to suggestions that a recognition model be learned off-line (Dayan et al., 1995; Hinton, Dayan, Frey, & Neal, 1995), which is exactly analogous to caching. The question as to whether a single model can support immediate, cached, recognition as well as on-line inversion, with the two competing for control and influence, has been raised (Dayan, 1999), but not satisfactorily addressed. Off-line learning need not only involve actual experience. Rather, samples produced ('dreamt') by the generative model can be used during times (notably sleeping, grooming and eating for rats, Diba & Buzsáki, 2007; Foster & Wilson, 2006; Lee & Wilson, 2002; Louie & Wilson, 2001; Skaggs & McNaughton, 1996) that active control is suppressed (Hinton et al., 1995; Sutton, 1991).

By comparison with model-based and model-free RL, other strands of work on automatization, notably associated with the SOAR, EPIC, CAPS and ACT-R architectures (Anderson et al., 2004; Just & Carpenter, 1992; Meyer & Kieras, 1997; Newell, 1990), adopt a rather different perspective, studying the properties and adaptation of complex policies for cognitively complex tasks, rather than the relatively simple policies that suffice for conditioning. However, there are some important links. For instance, for ACT-R (which we use as our running example), information from instructions can either be procedural, in the form of 'production' rules, or declarative, in the form of propositions.

Control based on a declarative representation of the problem is just like model-based control, with the same extreme demands on tree-based search. Control based on production rules is simple, as for a recognition model – being just an instantaneous pattern matching process.

In ACT-R, there is a sophisticated learning process (called knowledge compilation, (Anderson, 1982) or production compilation, (Taatgen & Anderson, 2002; Taatgen, Huss, Dickison, & Anderson, 2008)) that creates new procedures, for instance by filling specific declarative facts into the generic procedures or production rules that operate over those declarations, thereby creating specific productions. Analogous methods appear in other architectures such as *chunking* in SOAR (Newell, 1990; Newell & Rosenbloom, 1981). Of course, if the domain changes, then inconsistency can arise between the fast-changing declarative knowledge of the subjects and the filled-in details in the specific productions. The resulting apparent infelicity of behavior is the exact analogy of habitual control. Perhaps more troubling is that maintaining the logical (or indeed statistical) integrity of the deductions in the declarative system in the face of change in the world is an instance of the infamous frame problem for artificial intelligence (McCarthy & Hayes, 1969). Any invalidated deductions that are present in the declarative knowledge-base could give rise to habit-like behavior.

Note also that much use is made in ACT-R of a declarative, blackboard-like (Newell, 1994; van der Velde & de Kamps, 2006), fact-based working memory. This stores information about state that is implied by past inputs but is not present in the current sensation; something which is also required for both model-free and model-based control (formally, to create something akin to a Markov decision problem from a partially observable Markov decision problem). ACT-R's working memory also remembers partial results of calculations and subgoals; tree search for model-based RL requires a simple version of this too.

## 2.2. Interdependence versus independence

The strict separation between model-based (or declarative) and model-free (or procedural) methods is an idealization (Hammond, 1996), with at least four different interdependencies being discussed in the literature. First, we described model-free methods as using storage rather than calculation to work out the future value of present actions at current states. However, this storage is not best described as a simple look-up table mapping states to values, but rather as a mapping from an internal *representation* of the states to their values. Many different sorts of representations have been considered – indeed there is a whole field of unsupervised learning as a model of activity-dependent development (of which the paired generative and recognition models mentioned above is one example) that is concerned with the induction of different representations of sensory inputs based on their statistical structure (Rao et al., 2002). In the case of control, one idea has been that states should be represented in terms of their *successors* (Dayan, 1993), a representation that is then itself a form of model of the world (Sutton, 1995). Model-free control using such representations in the state-value or  $Q$  functions can automatically be endowed with some of the properties of model-based control.

Second, we described model-based evaluation as searching all the way to the leaves of the tree of states and actions by which point the outcomes will be clear. If this is hopelessly intractable, it is possible to search to some tractable depth, and then use the model-free state value function to substitute for the search that would lie below. This is completely conventional, for instance, for the case of playing board-like games such as chess (though it is not ubiquitous, for instance not featuring in the most recent work in

the game of GO; (Gelly & Silver, 2008)). This coupling has not been well explored in behavioral or neurobiological terms.

Third, in architectures such as ACT-R, one of the most important effects of a procedure (which we have identified more closely with model-free control) is to post a subgoal in working memory. Solving this subgoal then becomes the object of other effort. By itself, the subgoal is a declarative fact; it may be solved either by mechanisms operating directly over declarative knowledge about the domain, and thus be model-based, or procedurally, given a more advanced stage of knowledge compilation. Thus, just as we argued that model-based control should stealthily adopt model-free values, procedural control should stealthily adopt declarative control where necessary to construct a whole policy. Reinforcement learning has embraced subgoals in the form of *options* (Sutton, Precup, & Singh, 1999), but currently lacks a range of sophisticated and powerful methods for inducing options automatically from the observed structure of a task.

Finally, one of the most important lessons from the explicitness of large-scale cognitive architectures is that model-based or declarative control is not a self-contained method of control, but rather depends richly on model-free, procedural, mechanisms for its calculations and instantiation. The procedural explicitness of ACT-R in this respect makes possible the seamless co-existence of declarative and procedural control, blurring the divide described above.

## 2.3. Interpreted versus compiled control

Even after the proceduralization of declarative rules in ACT-R, there is the possibility of further specialization, generalization and tuning of the collection of productions to achieve superior performance (the 'autonomous' stage of performance in the terms of Fitts (1964)). Dayan (2007) considered the two extremes in the context of a neural proposal for the implementation of rules to solve problems such as the conditional one-back task (Frank, Loughry, & O'Reilly, 2001) that are much simpler than columnar arithmetic or air-traffic control (Anderson & Lebiere, 1998) but might nevertheless pose severe difficulties for non-human primates.

Dayan (2007) noted that such problems can either be solved with simple rules that are only relevant in very particular circumstances and therefore need substantial checking to ascertain that they are valid, or with complex rules that operate over wide domains, and thus require less checking. He suggested a uniform, neurally-inspired, architecture for combining simple and complex rules, with an associative recall for the rules coupled to an explicit process for testing the match between the preconditions of the rules and the current working memory (and sensory input). However, unlike Anderson (1982), this study did not articulate an explicit method which actually carries out the ultimate automatization of generating the complex rules.

One can make the analogy between these two forms of solution and interpreted and compiled programming languages. Rule matching and checking associated with the simple rules are the on-line operations necessary for interpretation. The creation of the complex, multi-functional rules that operate autonomously, without the need for repeated checking, is another form of compilation and automatization (albeit different from the knowledge or procedure compilation involved in turning declarative knowledge into procedures).

There is a hierarchy of levels joining paradigmatic examples of interpretation and compilation. Rather speculatively, Dayan (2007) suggested that there is a parallel between this sort of hierarchy and that apparent in the structure of the representation in the brain of sensory (notably visual Felleman & Essen, 1991) inputs, and, furthermore, that the same sort of unsupervised learning procedures that are believed to create the latter in sensory cortical areas (which we mentioned above) might create the former in premotor and prefrontal cortical areas.

#### 2.4. Prior-bound versus data-bound control

From a Bayesian perspective, when solving a new control problem, it is necessary to combine prior expectations with observations. These prior expectations can come in many different forms – everything from the overall statistics of rewards or punishments or the strength or reliability of the degree of control that a subject might be able to exert over the environment (Huys, 2007; Huys & Dayan, 2008). These two factors might also interact, for instance with the possible asymmetry that, in natural environments, rewards are typically rare and caused by a subject's own actions; whereas punishments (or at least threats) are typically common, and caused by the actions of others (Dayan & Huys, 2009).

Even rather generic priors can have a powerful and even pernicious effect in control problems because of the tradeoff between exploration and exploitation. In domains that are partially unknown, subjects have to choose whether to devote effort to *exploring* the consequences of actions that they do not know versus *exploiting* their existing knowledge and making the best choice in the light of that. Exploration is only worthwhile if there is a good chance of (a) finding an action with a better outcome than the existing actions; and (b) having this outcome be a reliable consequence of this action. Both of these are dependent on the current beliefs about the world. Thus, if, for instance, the agent starts with complete ignorance but a pessimistic view of the reward and control characteristics of the environment, then its willingness to explore will be limited, and thus it might not find out that more optimism would have been called for. Huys (2007) and Huys and Dayan (2008) suggested this as an account of learned helplessness (Maier & Seligman, 1976), a prominent animal model of depression, in which exposure to one environment in which they have no control over the termination of an unpleasant event causes subjects to fail to attempt to exert control over other environments.

We have argued that model-based control, because of its greater statistical efficiency, dominates in the face of little evidence. Thus, we would expect it to be more affected by priors than model-free control, which is data-bound because of the substantial experience necessary for it to dominate. However, the exploration–exploitation trade-off described above implies that priors can affect model-free, habitual, control too. Since biases about the environment manipulate the way that goal-directed control performs sampling, they can limit the experience that the model-free controller uses for learning. This can obviously affect its ultimate choices.

There is another set of what might be seen as evolutionarily-specified priors (Bolles, 1970; Breland & Breland, 1961; Dayan, Niv, Seymour, & Daw, 2006; Dayan & Seymour, 2008; Kahneman & Frederick, 2007). In Pavlovian conditioning, action-independent predictions about future outcomes (such as future rewards) lead to behavioral choices (such as approach) that are not only irrelevant to the delivery of those outcomes (by experimental design), but can actually be detrimental to that delivery (as in omission schedules or negative automaintenance Williams & Williams, 1969). These powerful prior policy biases are generally appropriate in natural environments; but their unsuitability in experimentally-controlled cases is revealing about the underlying mechanisms of control. Although there is some evidence about the neural structures involved in these Pavlovian choices (Reynolds & Berridge, 2001, 2002, 2008), it is only a hypothesis, based on the fact that model-based mechanisms can actually predict the inappropriateness of the consequences of these Pavlovian responses, that they are more parasitic on habitual than goal-directed control (Dayan et al., 2006).

#### 2.5. Instructed versus learned control

One main difference between architectures like ACT-R and both model-free and model-based reinforcement learning is that the former emphasises, or at least allows, human verbal instruction whereas the latter emphasises learning. One of the main forms of verbal instruction is a set of declarative facts that specify the structure of a control domain, possibly including collections of goals and subgoals, but without providing the inverse or recognition policy appropriate to that domain. In fact, the ideal case of providing 'pure' information about goals without any corruption from performance in a domain for reward (that could also train a model-free system) is hard to conceive without some form of verbal instruction.

However, verbal instruction is not confined to the goal-directed system, since precise recipes or procedures are also frequently provided, or some combination of procedures together with collections of goals and sub-goals. Further, model-based and model-free learning can also be structured, for instance through the *shaping* of competence, building complex functionality in a step-by-step manner (Krueger & Dayan, 2009; Skinner, 1938). The form of instruction that leads to the fastest learning, and most flexible and robust ultimate skills is under active investigation (Taatgen et al., 2008); the ongoing and highly charged debates about the most effective methods of teaching well illustrates our ignorance in this domain.

Finally, in all cases, turning verbal instructions into the precise internal representations of either declarative facts or productions is itself an important problem, whose solution lies in the underplumbed complexities of neural processing of language.

#### 2.6. Prefrontal cortex versus basal ganglia substrates

The concrete boundary in the animal conditioning literature between model-based and model-free control has made it possible to start elucidating the neural substrates of each. Evidence from lesion studies in rats (Balleine, 2005; Killcross & Coutureau, 2003), confirmed to a surprising degree in human functional magnetic resonance imaging (fMRI; Tanaka, Balleine, & O'Doherty, 2008; Valentin, Dickinson, & O'Doherty, 2007), implicates regions of the prefrontal cortex and their connections to dorsomedial striatum in goal-directed control, and the dorsolateral striatum in habitual control. There is also anatomical (Haber, Fudge, & McFarland, 2000; Joel & Weiner, 2000) and lesion-based (Belin & Everitt, 2008) evidence that habits migrate along a ventral-dorsal axis of the striatum as they become more deeply embedded, suggesting that it will be desirable to make further subdivisions in the category of non-goal-directed control.

In rats, reversible lesions in the prelimbic and infralimbic areas of the medial prefrontal cortex are able to suppress goal-directed and habitual choices, respectively (Killcross & Coutureau, 2003), which is one of the main sources of evidence that these systems operate in parallel rather than serially. Furthermore, the influence of Pavlovian predictions on the selection of actions appears to involve the nucleus accumbens and ventral striatum in both rats and humans (Balleine, 2005; Bray, Rangel, Shimojo, Balleine, & O'Doherty, 2008; Reynolds & Berridge, 2001, 2002; Talmi, Seymour, Dayan, & Dolan, 2008).

There is also evidence that the orbitofrontal cortex (O'Doherty, 2007; Rolls & Grabenhorst, 2008) and the basolateral nucleus of the amygdala (Balleine & Killcross, 2006) are involved in representing model-based values of outcomes, and the central nucleus of the amygdala in representing model-free values (Balleine & Killcross, 2006). The neuromodulator dopamine has many of the correct properties to direct learning of the model-free system (Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Morris, Nevet, Arkadir,





