Cognitive and neural foundations of goal-directed decision making:
An integrative framework

Matthew M. Botvinick, Steven Ibara and Janani Prabhakar
Princeton Neuroscience Institute and Department of Psychology
Princeton University

October 14, 2009

Running head: Goal-directed decision making

Word count: 20,460

Corresponding author:

Matthew Botvinick
Princeton University
Department of Psychology
Green Hall
Princeton, NJ 08540
(609) 258-1280, fax (609) 258-1280
matthewb@princeton.edu
Abstract

Recent work has given rise to the view that reward-based decision making is governed by two key controllers: a habit system, which stores stimulus-response associations shaped by past reward, and a goal-oriented system that selects actions based on their anticipated outcomes. The current literature provides a rich body of computational theory addressing habit formation, centering on temporal-difference learning mechanisms. Less progress has been made toward formalizing the processes involved in goal-directed decision making. We draw on recent work in cognitive neuroscience, animal learning, cognitive and developmental psychology and machine learning, to propose a new theory of goal-directed decision making. Our basic proposal is that the brain, within an identifiable network of cortical and subcortical structures, implements a generative model of action: a representation of the interdependencies among states, actions, outcomes, and rewards. Goal-directed decision making is understood as involving a set of procedures for querying this generative model, in order to extract plans for action. We focus on one particularly interesting and powerful form of query, which involves reasoning from the assumption of future reward. The proposed account is formalized in a set of simulations addressing benchmark empirical phenomena, and tested through two human behavioral studies. We discuss the relationship between the proposed framework and other models of decision making, including recent models of perceptual decision making, to which our theory bears an intimate but surprising connection.
Since the earliest days of both psychology and neuroscience, investigators interested in decision making and the control of behavior have recognized a fundamental distinction between habitual action and goal-directed or purposive action. Although the roots of this opposition clearly lie in commonsense notions from folk psychology, some more precise and operational distinctions arose from the classic debates of the first half of the 20th century. There, Hull (1943), Spence (1956) and others characterized action selection as driven primarily by immediate associations from internal and environmental states to responses, while others including Tolman (1932) and McDougall (1923) portrayed it as arising from a process of prospective planning, involving the anticipation, evaluation and comparison of action outcomes.

Tolman and other purposivists supported their view by pointing to specific empirical results from research with animals. Two findings that received particular emphasis were latent learning and detour behavior. In Blodgett’s (1929) classic demonstration of latent learning, rats ran a compound T-maze, as shown in Figure 1 (left), until they reached the box labeled ‘exit.’ After several sessions, a food reward was placed in the exit box. After the animals discovered this change, there was an immediate reduction in the frequency of entrances into blind alleys. Animals suddenly took a much more direct path to the exit box than they had previously.

In detour behavior, as described by Tolman and Honzik (1930), rats ran a maze configured as in Figure 1 (right). When the most direct route (path 1) was blocked by a barrier at position A, the animals tended to opt for the shortest of the remaining paths (path 2). However, when the block was placed at point B, animals took the third path.

Tolman (1932; 1948) offered an explanation of latent learning, detour behavior, and many related phenomena, based on the idea of a cognitive map: a knowledge structure representing the causal relations between actions and outcomes. According to Tolman (1932; 1949) purposive behavior involves the coordination of such causal knowledge with incentive knowledge, knowledge of the reward values associated with individual world states. Integration of these two forms of knowledge allows the selection of actions judged most likely to bring about preferred outcomes. This general framing of goal-directed decision making has endured, in very much the same form, to the present day. For example, according to Balleine and Dickinson (1998b), “the actions of animals are determined by the integration of two sources of information: (1) beliefs concerning the consequences of actions acquired during training and (2) the current desirability of the outcome of these actions” (p. 69).

Recent research, carrying forward the tradition represented by Tolman, has introduced an important additional experimental approach to diagnosing goal-directed or purposive behavior, based on outcome devaluation (Adams & Dickinson, 1981; Balleine, 2005; Balleine & Dickinson, 1998c; Colwill & Rescorla, 1985b; Klossek, Russell, & Dickinson, 2008). Here, an animal first learns to perform actions that yield specific rewards, for example, learning to pull a chain that yields one kind of food and to press a lever that yields another. The appeal or reward value of one of the outcomes is then reduced, for example, with food rewards by either by allowing the animal to eat its fill of one food (the specific satiety procedure; Balleine & Dickinson, 1998d; Colwill & Rescorla, 1985a) or by pairing that food with an aversive event such as toxin-induced illness (conditioned aversion; Adams, 1982; Adams & Dickinson, 1981; Colwill & Rescorla, 1985a; Colwill & Rescorla, 1988). Under appropriate circumstances, this intervention results in a
rapid shift in behavior away from the actions associated with the devalued outcome. Such a shift is interpreted as reflecting goal-directed behavior in the sense that it implies an integration of action-outcome knowledge with representations of outcome reward value.

In Pursuit of a Computational Account

Partly as a result of research utilizing the outcome devaluation paradigm, the early view of habit and goal-directedness as mutually exclusive accounts of behavior has given way to a more inclusive multiple-systems account, under which habitual and goal-directed control coexist as complementary mechanisms for action selection (Daw, Niv, & Dayan, 2005; Dickinson, 1985; Dickinson & Balleine, 1993). This perspective justifies the study of each form of action control in its own right, and of course large literatures have developed concerning both habitual stimulus-response-based action selection (Frank & Claus, 2006; Wood & Neal, 2007; Yin & Knowlton, 2006) and prospective planning-based control (Bargh, Green, & Fitzsimons, 2008; Bekkering, Wohlschlager, & Gattis, 2000; Gergely & Csibra, 2003; Gollwitzer & Moskowitz, 1996).

However, if these two literatures are considered side by side, a striking asymmetry becomes evident, which relates to computational and mechanistic precision. Research on habitual, stimulus-response behavior has, over the past decade, crystallized around an increasingly explicit set of computational ideas, originating in the field of computational reinforcement learning (Sutton & Barto, 1998). These ideas have not only provided a context for interpreting and predicting patterns of behavior (Barto & Sutton, 1981; Sutton & Barto, 1990; Wickens, Kotter, & Houk, 1995); they have also enabled new and detailed insights into the functional contributions of specific brain structures, including the striatum and the midbrain dopaminergic system (Barto, 1995; Houk, Adams, & Barto, 1995; Joel, Niv, & Ruppin, 2002; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). In contrast, research on goal-directed behavior, for all its sophistication, has not developed a comparably mature computational core. Of course, cognitive psychology has generated detailed and often fully implemented computational accounts of planning and problem solving (Anderson, 2004; Laird, Newell, & Rosenbloom, 1987). Such models make contact with fundamental issues in goal-directed action selection, such as the ubiquitous role of reasoning from ends to means (Newell & Simon, 1971). However, because such models generally allow no central role for incentive value, they address only one half of the full computational problem posed by goal-directed behavior, which, as we have noted, is defined by an integration between causal and reward knowledge.

Models that do admit a central role for reward or utility can of course be found in the literature on economic and neuroeconomic decision making (see, e.g. Glimcher, Camerer, Fehr, & Poldrack, 2009). However, the focus in such work has tended to be on characterizing observed or normative decision outcomes rather than on specifying the algorithms through which decisions are made. And in cases where attention has been directed toward the algorithmic (or implementational) level, the focus has been primarily on single-step decision making, and the distinction between habitual and goal-directed decision has not always been remarked. What is needed, essentially, is a framework that integrates the economic and neuroeconomic focus on reward with the focus, in traditional cognitive psychological models, on (potentially complex) causal knowledge.

Spreading Activation Models of Goal-Directed Decision Making

One group of computational models that does begin to address the integration between causal and incentive knowledge has been put forth by Schmajuk and colleagues (Schmajuk & Thieme,
Botvinick and colleagues  

Goal-Directed Decision Making

1992; Voicu & Schmajuk, 2002), and more recently by Hasselmo and colleagues (Hasselmo, 2005; Koene & Hasselmo, 2005). Both sets of models assume a network of simple neuron-like processing elements representing environmental states or locations, which plays the role of the cognitive map. In Schmajuk’s models, which address spatial navigation tasks, inputs representing incentive value activate rewarded locations, and activation spreads from these locations to adjacent ones until the frontier of activation reaches the agent’s current location. This results in an activation map, from which actions can be selected through a hill-climbing procedure (for related work see Bugmann, Taylor, & Denham, 1995; Gaussier, Revel, Banquet, & Babeau, 2002; Girard, Filliat, Meyer, Berthoz, & Guillot, 2005; Martinet, Passot, Fouque, Meyer, & Arleo, 2008; Muller, Stead, & Pach, 1996; Reid & Staddon, 1998). Hasselmo’s models (Hasselmo, 2005; Koene & Hasselmo, 2005) follow this same general approach, but allow activation also to spread ‘forward’ from the agent’s initial state. These models also explicitly represent actions and action-outcome relationships, permitting the models to be applied beyond the domain of spatial navigation.

These spreading activation models take an important step in the needed direction, providing an initial sense of how something like a cognitive map might be implemented in neural hardware, along with a coarse account of how reward information might be integrated with the knowledge represented in such a map to guide decision making. However, there is clearly room for further development. The range of scenarios that these previous models can address is limited. For example, it is not clear whether it would be possible to address situations involving multiple competing sources of reward or cost, or situations in which action-outcome relationships are probabilistic. At a more basic level, although Hasselmo (2005) does draw some links to reinforcement learning, neither of the modeling paradigms we have mentioned is founded on a formal, normative analysis of the task to be solved. While beginning with such an analysis is certainly not a scientific requirement, it is a strategy whose power has been demonstrated in many recent models of decision making of other kinds (e.g., Anderson, 1990; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Geisler, 2003; Niv, Joel, & Dayan, 2006).

Goal-Directed Decision Making as Model-Based Reinforcement Learning

A potential foundation for a more formally explicit approach was pointed out by Daw, Niv and Dayan (2005) (see also Dayan & Niv, 2008), who suggested that goal-directed decision making should be viewed as a form of model-based reinforcement learning. The ‘model’ referred to in this phrase comes in two parts: a state-transition function, which maps from situation-action pairs to outcomes, and a reward function, which attaches a scalar reward value to each world state. Model-based reinforcement learning refers to the project of inferring an optimal (reward-maximizing) policy, or mapping from states to actions, given this two-part model (Sutton & Barto, 1998).

In a way, framing goal-directed decision making in these terms simply recapitulates the classical account, with the state-transition function playing the role of the cognitive map, and incentive knowledge encoded in the reward function. However, by drawing the parallel to model-based reinforcement learning, Daw, Niv and Dayan (2005) signaled the relevance of established formalisms in machine learning and operations research. As it happens, the problem of finding an optimal policy based on state-transition and reward functions has been extensively studied in these fields, under the rubric Markov decision processes or stochastic optimal control (Bertsekas & Tsitsiklis, 1996; Puterman, 2005; Sutton & Barto, 1998).

Given the success of research investigating the potential relevance of reinforcement learning methods to habit-based control in animals and humans, it is inviting to consider whether model-
based reinforcement learning might be of use in understanding goal-directed decision making. The literature on Markov decision processes provides a range of techniques to consider, including backward induction, policy iteration, value iteration, and methods based on linear programming (Bertsekas & Tsitsiklis, 1996; Puterman, 2005; Sutton & Barto, 1998). In the present article, we examine the applicability of a relatively recent approach, which frames the problem in terms of probabilistic inference.

**Goal-Directed Decision Making and Probabilistic Inference**

In introducing this perspective, we begin by returning to Tolman’s cognitive map. In recent work, Gopnik and colleagues have proposed an updated view of the cognitive map (Gopnik et al., 2004; Gopnik & Schulz, 2007). Based on empirical data on causal reasoning in both children and adults, they suggest that mental representations of the world’s causal structure can be understood as networks of probabilistic relations, well described by the directed graphical models or Bayesian networks that have recently seen extensive application in statistics (Bishop, 2006; Koller & Friedman, 2009; Pearl, 1988). Although the link to Tolman’s idea of a cognitive map is not always as explicit, a number of other researchers have also analyzed both human and animal causal representations in terms of graphical models (Blaisdell, Sawa, Leising, & Waldmann, 2006; Glymour, 2001; Sloman, 2005; Tenenbaum, Griffiths, & Niyogi, 2007). In addition to the power of this approach to account for empirical data, it is appealing because it frames the representation of causal structure in terms that have been recently applied, in very much the same form, to many other domains of human and animal information processing, including perception (Gold & Shadlen, 2001; Kersten & Yuille, 2003; Knill & Pouget, 2004), motor control (Carpenter & Williams, 1981; Kording & Wolpert, 2006), language (Chater & Manning, 2006; Xu & Tenenbaum, 2007), memory (Hemmer & Steyvers, 2009), and conceptual knowledge (Chater & Oaksford, 2008; Griffiths, Steyvers, & Tenenbaum, 2007). Furthermore, the same terms have also been applied to neural information processing, which has been modeled, at both the single-neuron and network levels, in terms of Bayesian inference within probabilistic graphs (Beck & Pouget, 2007; Litvak & Ullman, 2009; Ma, Beck, Latham, & Pouget, 2006; Pouget, Dayan, & Zemel, 2003; Rao, 2006). Conceptualizing the cognitive map in probabilistic terms thus aligns this construct with a set of emerging domain-general representational and information-processing principles. Given the central role of the cognitive map in goal-directed decision making, this re-conceptualization suggests the appealing possibility of accounting for goal-directed decision making in the same domain-general terms.

Of course, recasting the cognitive map as a causal Bayesian network does not, in itself, yield a theory of goal-directed behavior. Still lacking is an account of how the causal network might be leveraged in order to select actions. More specifically, an account is needed of how reward information is represented, and how this information is integrated with the information in the causal model in order to arrive at adaptive policies.

Fortuitously, recent work in machine learning has begun to address precisely these questions. Alongside classical techniques for solving Markov decision problems, such as policy and value iteration, a number of approaches have arisen that are based on inference within probabilistic graphs (Attias, 2003; Botvinick & An, 2009; Cooper, 1988; Dayan & Hinton, 1997; Hoffman, de Freitas, Doucet, & Peters, 2009; Shachter & Peot, 1992; Toussaint & Storkey, 2006; Verma & Rao, 2006b). Such work assumes a causal model of the kind described by Gopnik and colleagues (2004), but embeds this within a larger probabilistic network, which also contains representations of actions and rewards. This larger network provides the substrate for inference procedures that solve the planning problem, yielding rational policies for action.
We believe that these advances within machine learning may provide a useful foundation for developing theories of goal-directed behavior in humans and other animals, in much the same way that temporal-difference learning algorithms have provided a foundation for theories of habitual behavior. It is this possibility that we examine in the present paper.

Overview

In what follows, we develop a computational framework for understanding goal-directed decision making, framed in terms of structured, model-based probabilistic inference. The work is inspired by, and draws from, previous efforts within machine learning to solve Markov decision problems through inference within probabilistic graphical models (Attias, 2003; Cooper, 1988; Shachter & Peot, 1992; Toussaint & Storkey, 2006; Verma & Rao, 2006b), although, as we shall discuss, it is not formally identical to any previous approach. More importantly, unlike previous machine learning work, the account we offer here is explicitly proposed as a theory of decision making in animals and humans. With this in mind, we draw parallels — some firm, others tentative — between components of our model and specific neural structures, most prominently regions within the prefrontal cortex. We also apply the computational framework, through simulation, to qualitative behavioral findings from research on goal-directed decision making. Finally, we present results from two new behavioral experiments, testing specific predictions arising from the computational framework.

A Computational Framework for Goal-Directed Decision Making

We begin with the basic assumption that there are four general domains of representation that are relevant to goal-directed decision making. These encompass, respectively, (1) policy representations, (2) action representations, (3) representations of environmental and internal state, and (4) representations of reward or utility. As detailed in what follows, we propose that these domains of representation map onto distinct, though overlapping, sets of neuroanatomical regions. Equally importantly, on a computational level, the four domains interface with one another in a specific way, forming a network that supports rational decision making. Over the following four sections, our strategy will be to build up this architecture piece by piece, situating new components within it as they are introduced.

Policy Representations and the Dorsal Prefrontal Cortex

Investigations of the neural substrates of goal-directed decision making provide convergent evidence for the central involvement of dorsal prefrontal cortex. Lesion studies in rodents have shown a reduction in sensitivity to outcome devaluation following lesions to prelimbic cortex (Balleine & Dickinson, 1998a; Corbit & Balleine, 2003; Killcross & Coutureau, 2003) (however see Ostlund & Balleine, 2005, discussed below), indicating a loss of one hallmark of goal-directed behavior. The prelimbic cortex has been judged to represent an approximate homologue to the dorsolateral prefrontal cortex (DLPFC) in primates, including humans (Fuster, 1997; Kesner, 2000), a region that has been extensively implicated in planning and goal-direction (Anderson, Albert, & Fincham, 2005; Duncan, Emslie, Williams, Johnson, & Freer, 1996; Goel & Grafman, 1995; Lengfelder & Gollwitzer, 2001; Miller & Cohen, 2001; Shallice, 1982; Shallice & Burgess, 1991; Tanji & Hoshi, 2008; Tanji, Shima, & Mushiake, 2007; Unterrainer & Owen, 2006). Single-unit recording studies in primate DLPFC, and complementary functional neuroimaging studies, have indicated that one important function of the region may be to represent task sets or ‘rules’ (Asaad, Rainer, & Miller, 2000; Bunge &
Botvinick and colleagues Goal-Directed Decision Making

Wallis, 2007; Sakai, 2008; Wallis, Anderson, & Miller, 2001; White & Wise, 1999). Importantly, the content of such rules is typically understood as establishing a set of relationships between stimuli and responses (Bunge, 2004), something that comes very close to a policy, as we defined the term earlier. The concept of policy representation also applies to work addressing how DLPFC activation exerts an impact on overt behavior. According to the guided activation theory of Miller and Cohen (2001), a critical function of the DLPFC is to bias the flow of neural activation in pathways between stimulus and response representations, supporting transmission along task-relevant pathways. Here, through its effect on processing in other regions of the brain, the DLPFC establishes a mapping from stimuli to responses, i.e., a (partial and temporary) policy.

Note that one can rephrase the guided activation theory in probabilistic terms by expressing the relationship between environmental states \( (s) \) and responses \( (a) \) as a conditional probability distribution, \( p(a|s) \), and then adding that this distribution depends on, or is parameterized by, the policy representation \( (\pi) \) currently active in DLPFC: \( p(a|s,\pi) \). This relationship is diagrammed in Figure 2A, in the form of a directed graphical model. The node \( s \) represents a random variable with one discrete value for each state of the environment and organism, as represented across brain regions involved in perception, working memory and homeostasis. This node is filled in to indicate that its value is known or observed by the decision-maker; the current state is a ‘given’ in the action-selection problem.

FIGURE 2 AROUND HERE

The node \( a \) represents a random variable with one discrete value for each possible action or response. The arrow from \( s \) to \( a \) indicates that the value of \( a \) depends on \( s \). As indicated by the second arrow in the figure, \( a \) also depends on a third (discrete, multinomial) variable, which represents the policy currently selected in DLPFC.

The network in Figure 2A gives us a simple way of representing the statistical or causal relationships among states, actions and policies. Our objective is to specify how knowledge of these relationships might be used to support the selection of optimal policies. Before taking the next step in this direction, however, a caveat is needed regarding our focus on DLPFC as a locus of policy representation. For, in addition to the evidence implicating this region, there is also data suggesting that policy, task set, or rule representations may also reside in other portions of the frontal lobe, including premotor cortex (Wallis & Miller, 2003), ventrolateral prefrontal cortex (Bunge, 2004; Bunge et al., 2005), pre-supplementary area (Dosenbach et al., 2006; Rushworth, Walton, Kennerley, & Bannerman, 2004), and the frontal pole (Sakai & Passingham, 2003). In addition, the data concerning the role of rodent prelimbic cortex in goal-directed decision making is not without nuance (e.g., Ostlund & Balleine, 2005). In view of these points, the most conservative approach is to stipulate that policies are represented, in a distributed fashion, over a system of frontal regions. Having said this, it remains true the preponderance of evidence points to the DLPFC as an important hub in this policy system, and in what follows we will continue to identify the role of policy representation with this area of the cortex.

Representing Future States and the State-Transition Function

The processing architecture that we have constructed so far includes a representation of the present state. However, a critical feature of goal-directed decision making is that it also takes explicit account of future states, i.e., projected action outcomes. We can integrate such states into the architecture by inserting a second state node, as shown in Figure 2B. This node
receives arrows from the original action and current-state nodes. As before, these arrows represent causal dependencies. Together, they indicate that the probability distribution over values of the new state variable, $s'$, depends jointly on the initial state $s$ and the distribution over actions, $a$.

Note that the resulting conditional probability distribution (CPD) $p(s'|s,a)$ is the state-transition function that forms one part of the basis for model-based reinforcement learning. It is also functionally equivalent to the cognitive or causal map described by theorists from Tolman (1932) to Gopnik and colleagues (see also Colwill & Rescorla, 1990; 2004).

It is clear that the brains of higher animals must contain representations of projected future states, as well as their dependencies on earlier states and actions (Atance & O’Neill, 2001; Gopnik & Schulz, 2007; Schutz-Bosbach & Prinz, 2007). However, despite considerable research, the neuroanatomical site of such representations is only beginning to emerge. Early studies of spatial navigation in rodents led to the idea that cognitive map representations might reside in the hippocampus (O’Keefe & Nadel, 1978), and, consistent with this, recent research suggests that hippocampal place cells may represent projected future locations, consistent with a role in route planning (Diba & Buzsaki, 2007; Johnson & Redish, 2007; Johnson, van der Meer, & Redish, 2008). Convergent neuropsychological research in humans indicates that medial temporal lobe structures play a critical role in allowing visualization of future events (Hassabis, Kumaran, Vann, & Maguire, 2007). Research with rodents provides direct evidence for the involvement of medial temporal lobe structures (entorhinal cortex, if not hippocampus) in the representation of action-outcome contingencies during instrumental learning (Corbit, Ostlund, & Balleine, 2002).

In recent work, Schacter, Addis and Buckner (2007) and Buckner and Carroll (2006) have portrayed the medial temporal lobe as part of a larger system supporting the imaginative projection of future events, including action outcomes. Also included in this putative network are regions within medial and lateral parietal cortex (see also Hamilton & Grafton, 2008), lateral temporal cortex and medial frontal cortex (see also Matsumoto, 2004; Matsumoto, Suzuki, & Tanaka, 2003; Tanaka, Balleine, & O’Doherty, 2008). Still other work has suggested that the DLPFC may play a role in representing projected action outcomes, including both final ‘goal’ states and intermediate ‘means’ states (Fuster, 1997; Mushiake, Saito, Sakamoto, Itoyama, & Tanji, 2006; Saito, Mushiake, Sakamoto, Itoyama, & Tanji, 2005), and a recent study by Hamilton and Grafton (2008) suggests that the right inferior frontal cortex may be also be involved in representing action outcomes.

At the subcortical level, there is strong evidence for the involvement of specific basal ganglia structures in the representation of action-outcome contingencies. Research in rats has shown that damage to or inactivation of the dorsomedial striatum impairs sensitivity to outcome devaluation and changes in instrumental contingency (Balleine, 2005; Yin, Knowlton, & Balleine, 2005; Yin, Ostlund, Knowlton, & Balleine, 2005). This fits well with research implicating the caudate nucleus, the primate homologue of the dorsomedial striatum, in planning and problem solving (Monchi, Petrides, Strafella, Worsley, & Doyon, 2006; Unterrainer & Owen, 2006). It also squares with recent human neuroimaging work by Tanaka, Balleine and O’Doherty (2008), in which caudate activation was observed to correlate with the degree to which salient (reward) outcomes were contingent on participants’ actions.

Taken together, existing research suggests that a diverse network of regions contribute to the projection of action outcomes. In the model we are developing, we shall treat this network as a
unit, referring to it as the state projection system in order to distinguish it from the set of brain regions involved in representing current environmental and homeostatic states.

Incentive Value, Orbitofrontal Cortex and Basolateral Amygdala

As established earlier, true goal-directed decision making rests not only on causal information — a state-transition function — but also on information about incentive values attached to outcome states: a reward function. The relationship between states and reward values can be represented within the structure we have been building by adding one further node, which we shall refer to as a reward node, and label $u$ (Figure 2C). As explained further below, this node, like the others, represents a random variable, in this case a variable encoding reward magnitude. The arrow running to it indicates that the value of this reward variable depends on which outcome state (or probability distribution across outcome states) is predicted.

As with the other variables we have built into the model, we intend reward variables to capture, in the simplest possible way, the representational role of a specific set of brain regions. In this case, the relevant regions include, most prominently, the orbitofrontal cortex and the basolateral amygdala. The orbitofrontal cortex (OFC) has been extensively implicated, across species, in the representation of the incentive value of stimuli, including anticipatory coding for the value of action outcomes (Kringelbach, 2005; Montague & Berns, 2002; Padoa-Schioppa & Assad, 2006; Rolls, 2004, 2006). This function has been linked, in part, to a role in goal-directed decision making (Frank & Claus, 2006; Roberts, 2006; Rolls, 1996; Schoenbaum & Setlow, 2001; Schultz, Tremblay, & Hollerman, 2000; Wallis, 2007). Such a role is indicated, for example, by disruptions in instrumental choice behavior following damage to OFC (Bechara, Damasio, Damasio, & Anderson, 1994; Bechara, Damasio, & Damasio, 1999; Fellows & Farah, 2003; Rolls, 1996, 1999), including a reduced sensitivity to devaluation (Pickens, Saddoris, Gallagher, & Holland, 2005). It is also reflected in elevated activity in OFC during decision making relating to the incentive value of outcomes (Arana et al., 2003; Plassman, O'Doherty, & Rangel, 2007), as well as in reductions in OFC responses to appetitive stimuli after devaluation (Gottfried, O'Doherty, & Dolan, 2003; Valentín, Dickinson, & O'Doherty, 2007). Lesions to OFC in monkeys have been found to reduce devaluation sensitivity (Izquierdo, Suda, & Murray, 2004), a hallmark of goal-directed decision making (though see Ostlund & Balleine, 2007 for some interesting contrary findings in rats).

Despite important differences in function, the basolateral amygdala (BLA) has also been extensively implicated in the representation of incentive value of stimuli, including action outcomes, and in the guidance of goal-directed behavior (Balleine, 2005; Baxter & Murray, 2002; Holland & Gallagher). As in OFC, amygdala engagement scales with the incentive value of stimuli (Arana et al., 2003) and can be reduced through outcome devaluation (Gottfried et al., 2003; LaBar et al., 2001). Furthermore, BLA lesions disrupt devaluation sensitivity (Balleine, Killcross, & Dickinson, 2003; Corbit & Balleine, 2005). Interestingly, there is evidence that devaluation sensitivity may depend on a functional interaction between BLA and OFC (Baxter, Parker, Lindner, Izquierdo, & Murray, 2000), a point that lends some face validity to our treatment of these two structures as part of an integrated system for reward representation (Cavada, Company, Tejedor, Cruz-Rizzolo, & Reinoso-Suarez, 2000; Schoenbaum, Setlow, Saddoris, & Gallagher, 2003).

Having introduced the reward system into our developing framework, more must now be said about the specific way in which reward values will be represented. Given that the $u$ node represents reward magnitude, it might seem most natural to treat it as a continuous variable, representing rewards as scalar values. This is, indeed, the traditional computational approach
(see, e.g., Attias, 2003). However, we shall adopt a different method for representing rewards. In work on decision analysis, Cooper (1988) introduced the idea of representing reward using a Bernoulli random variable, i.e., a variable with two discrete states, ‘zero’ (or ‘false’) and ‘one’ (or ‘true’). Continuously-valued reward magnitudes can be represented as probabilities over the variable’s two states. Using our variable label $u$, a large positive reward would be indicated by a value of $p(u)$ (i.e., $p(u = 1)$) close to one, a large negative reward (punishment) by a value close to zero.

Although the choice of a binary rather than continuous variable for reward representation may seem like a minor technical detail, it turns out to be an extremely useful step in transposing the goal-directed decision making into probabilistic terms, as work subsequent to Cooper (1988) has recognized (Botvinick & An, 2009; Dayan & Hinton, 1997; Furmston & Barber, 2009; Hoffman et al., 2009; Shachter & Peot, 1992; Toussaint & Storkey, 2006; Verma & Rao, 2006b), and as we shall directly demonstrate below.

### Representing Sequences of Action

The architecture that we have developed so far addresses only a single step of action. However, it can be easily extended to sequences of action. As shown in Figure 2D, this extension is accomplished by duplicating part of the existing structure, providing a series of state, action and reward nodes, one for each step of the action sequence.

In extending the architecture in this way, we also introduce one final new element: a binary variable representing the cumulative reward accrued over an action sequence ($u_G$). As explained further below (and see Tatman & Shachter, 1990), this final addition will play an important role in driving goal-directed decision making in sequential domains.

### Making Inferences

Probabilistic graphical models, like those we have been building, can be understood as representing the structure underlying the joint probability of a set of variables (Bishop, 2006; Jensen, 2001). Another way of expressing this is to say that the graph provides a generative model, which explains the joint probability distribution by revealing something about the processes that give rise to it. From this perspective, the structure illustrated in Figure 2C-D can be viewed as a generative model of action, which explains patterns of co-occurrence among policies, actions, states and rewards.

In addition to explaining observed patterns, generative models also directly support the making of inferences. One form of inference that will play an important role in what follows involves what we shall refer to as a query. Here, the value of one or more variables is assumed to be observed or known, and one asks about the posterior marginal probability distribution over another variable or set of variables, that is, the probability distribution given the observed variables.³

Let us draw a quick, and classic, illustration from Pearl (1988). Imagine you work an hour’s commute from home. One day, while at work, you receive a phone call from one of your neighbors, who just returned home to find that your house’s burglar alarm was ringing. On the drive home to check out the situation, you hear on the radio that a small earthquake occurred near where your home is located. “Oh,” you think, “it was probably the earthquake that set off the alarm.”
The basis for this conclusion can be captured in the simple causal model diagramed in Figure 3. As before, each node in this graph represents a variable (here, a binary true/false variable), and the arrows encode causal dependencies, whose quantitative form is stored in the CPD for each variable. The inference described above can be captured as a query applied to this structure. Specifically, the variables phonecall and radio are treated as observed (with value true), and one asks about the posterior distributions for earthquake and burglar.

**FIGURE 3 AROUND HERE**

An indication of the mathematical operations involved in answering this sort of query will be presented in the next section, in the context of goal-directed decision making. One important thing to notice for the present is the direction of inference in this example. Although the arrows in causal graphs always point from causes to effects, inference can flow in either direction. One can query with causes, in order to predict effects, but one can also query with effects, inferring causes, as in our example. The latter case, sometimes referred to as abductive inference, involves inverting the causal model using Bayes' law, as further detailed in the next section.

**Selecting Policies Through Queries**

The central proposal we wish to make is that goal-directed decision making, as it occurs in the brains of animals and humans, can be understood as involving queries over causal models of action. Further, rather than identifying a single form of query as relevant, we delineate a family of query types, each of which may underlie any particular instance of goal-directed decision making. As detailed in the following three sections, this family can be divided into three basic classes, according to the identity of the conditioning variables.

1. **Reasoning from Goals: State Queries**

According to a traditional view, goal-directed decision making starts with a pre-specified ‘goal’ state, and involves selecting actions by which that state can be realized (Newell & Simon, 1971; Unterrainer & Owen, 2006). Given its familiarity, this perspective provides a convenient starting point, though one that will ultimately prove surprisingly limited on its own.

Given a causal model like the one diagrammed in Figure 2C,D, planning to a predefined goal state can be accomplished by treating the initial and outcome states as observed variables, with the value of the outcome set to the goal state. The query then asks what the posterior probability is over all available policies. Such a state query is diagrammed in Figure 4 (top left).

**FIGURE 4 AROUND HERE**

As a concrete example, consider a rat at the center of an eight-arm radial maze, one arm of which is baited with food. Here, the initial state would be the hub location, the space of policies would include entry into each arm, and the possible outcome states (values of $s'$) could be labeled food and no food. The decision among available policies would be made by assuming the goal outcome — setting the value of $s'$ to food — and then inferring the conditional distribution over the competing policies.

In the language of the probability model, this query equates to asking for $p(\pi|s,s')$. Note that, using Bayes’ law, $p(\pi|s,s')$ can be rewritten as $p(s'|s) \cdot p(\pi) / p(s')$. If we assume (for the moment) that the prior probabilities over all policies are equal, and recognize the denominator...
\( p(s') \) as a normalization factor, it follows that \( p(\pi|s, s') \) must be proportional to \( p(s'|s, \pi) \). Given the relationships among the variables in the causal model, for each policy (value of \( \pi \)),

\[
p'(s'|s, \pi) = \sum_a p(s'|s, a) p(a|s, \pi)
\]

(1)

Since both factors on the right-hand side are given by the causal model, that model provides all the information needed to answer the query. Within the posterior distribution obtained, the policy with the highest conditional probability will necessarily be the one with the highest probability of bringing about the goal state.

In order to extend this form of query to the case of multi-step decision, consider the experiment performed by Mushiake et al. (2006). Here, monkeys viewed simple mazes, and moved a joystick to navigate from starting location highlighted in green to a red goal location. There were four available policies for each position in the maze: up, down, left and right, and the overall problem was to find the correct three-action sequence. Once again, the decision-making process could be modeled as involving a state query. Here, one state variable (aside from the initial state) would be treated as observed, its value set to the goal location (Figure 4, top right), and Bayes’ law would be used along with the conditional dependencies represented in the graph to infer the posterior probability distribution over the policy variables. Once again, the series of policies with the highest posteriors would indicate the appropriate sequence for accomplishing the goal.

Focusing on this experiment as an illustrative example allows us to postpone discussion of one or two computational subtleties. However, even without yet broaching these, it is possible to identify a number of serious limitations associated with the state query as a mechanism for goal-directed decision making. Indeed, given its close fit with intuitive notions of goal directedness, the state query, taken on its own, turns out to provide a surprisingly weak decision-making method.

One limitation is that, without further elaboration, the decision-making process does not take into account the relative appeal of different means or paths to the goal. Shortest-path behavior, like that shown in the detour experiments reviewed earlier (see Figure 1) clearly indicates that action programming takes preferences over means into account. Another problem is that reasoning from goal states hinders consideration of alternative outcomes. Consider the following scenario. A child walks into the kitchen to find that his mother has just baked a batch of cookies. He adopts the goal of eating one. However, he must decide whether to ask his mother for a cookie, or simply to snatch one when her back is turned. The latter approach, he reasons, is more likely to yield his goal. However, there is also a small chance that he’ll be caught and reprimanded, or even punished. Reasoning from the assumption that the goal will be fulfilled would necessarily mean ignoring these risks, and might thus result in a foolish decision.\(^5\)

Perhaps the deepest problem with the state-query approach is that beyond certain special contexts, including games and psychology experiments, explicit goals are rarely immediately given. The full problem in goal-directed decision making thus involves goal selection as an integral component (Oettingen & Gollwitzer, 2004). It is tempting to think of the decision-making process as divisible into two independent phases, the first involving goal selection and the next involving planning to the goal. However, as pointed out in expectancy-value theories (see Atkinson & Birch, 1978), adaptive goal selection must be guided by an assessment of candidate
goals’ feasibility, something that — in the general case — can only be evaluated through a consideration of available means.

Ultimately, these weaknesses of the state query as a method for goal-directed decision making stem from the same factor that limits traditional theories of planning: The decision-making procedure admits no role for reward information. In the next two sections, we consider two different forms of query, in which reward plays an important role.

2. Reasoning from Plans: Policy Queries

Figure 4 (middle) diagrams a different approach to inference-based decision making. Here, what is assumed is not an outcome state, but rather a policy. The question for inference is, given the chosen policy, what is the expected reward? To return to the radial maze example, the rat would here consider each policy (arm of the maze) in turn, choosing the policy associated with the largest expected reward.

Since, in our framework, reward is encoded as the posterior probability of a binary reward variable, applying this query to the graph in Figure 4 (middle left) amounts to asking for $p(s, a, u)$. Given the relationships among the variables in the graph, this can be factored as,

$$p(u|s, a) = \sum_{s', a} p(s')p(s'|s, a) p(a|s, \pi)$$

Once again, the terms on the right-hand side of this equation are provided directly by the causal model itself, making inference straightforward.

The same logic extends naturally to the context of multi-step decisions, such as the one involved in the maze task of Mushaike et al. (2006), introduced earlier. Here, each query involves stipulating a specific sequence of actions or policies, and evaluating its expected reward (Figure 4, middle right). Because, in the general case, rewards (and costs) may occur at any step in the sequence, this approach requires integrating rewards across time, something that is accomplished in our modeling framework by the cumulative reward variable $u_G$ (details provided below under Simulation Methods). In the multi-step case, the policy query thus seeks $p(u_G|s, \pi)$.

Because the policy query, as we have defined it, takes reward information directly into account, and projects outcomes rather than taking them as a foregone conclusion, it avoids the limitations associated with state queries, providing a more powerful and general decision-making method. The potential relevance of policy queries (or nearly equivalent procedures) has been recognized in recent theoretical work on animal decision making (see, e.g. Daw et al., 2005), and recent single-unit recording data in rodents provides compelling evidence for serial consideration of future actions and outcomes at behavioral choice-points (Johnson & Redish, 2007; Johnson et al., 2008).

Nevertheless, one thing that the policy query seems to give up, compared with the state query, is a direct role for outcomes in driving the decision-making process. Intuition strongly suggests — and previous models in both machine learning and cognitive psychology widely concur — that decision-making is at least partly based on reasoning backward from desirable outcomes (see, e.g., Duncan et al., 1996). This brings us to a third and particularly interesting form of inference-based decision procedure.
3. Reasoning from Reward: Reward Queries

Figure 4 (bottom) diagrams a third and final way to query the causal models of action that we constructed earlier. Here, rather than treating state or policy variables as observed, it is a reward variable that drives the query. Recalling that reward variables are binary in our framework, this form of query assumes $u = 1$ and asks for $p(\pi|s,u)$. Returning to the rat in the radial maze, the logic of this approach essentially involves the rat ‘assuming’ that its next action — not yet chosen — will yield maximal reward, and then ‘reasoning’ from that assumption to the optimal action: From reward to the food outcome, and from there to the appropriate policy.

The math describing this reward-based inference is quite similar to previous cases. The first step is to rewrite the quantities $p(\pi|s,u)$ using Bayes’ law (again assuming no pre-existing bias toward one policy over the others),

$$p(\pi|s,u) = p(\pi|s)$$

and then to factor the resulting expression into terms available from the causal model:

$$p(\pi|s,u) = \sum_{a,s'} p(a|s)p(s'|s,\pi)$$

In the more general multi-step setting, the reward query is implemented by treating $u_G$ as the observed variable, and inferring for each step $p(\pi|s,u_G)$ (Figure 4, middle right). However, here we run into an important technical issue. It turns out that simply computing the posterior distribution over each policy variable and then adopting the set of most probable values can backfire under some circumstances. This is because, in the setting of sequential decision-making, the optimal decision at any step depends on what actions are planned for later steps.

For illustration, consider the decision faced by the rat in the two-step T-maze illustrated in Figure 6 (see also Simulation 2A and Experiment 1 below). The numbers at the top of the figure indicate the reward value associated with items contained at the maze termini. Obviously, the optimal choice at the first decision point is to head left. However, this is only true if the animal’s plan at the next juncture, $S_2$, is to head left again. If the animal plans instead to head right if faced with decision point $S_2$, then the best choice at $S_1$ is actually to go right. The same is true if the animal has not yet made any decision about what to do at $S_2$ or $S_3$; if the animal is equally likely to head left or right at these points, then the best plan at $S_1$ is to go right.

This example illustrates the inherently recursive structure of multi-step decision making. A standard way of coping with this recursiveness in computational reinforcement learning is by making the decision-making procedure iterative: within a single decision-making episode the policy is updated repeatedly, with the interim result on each iteration feeding into the next update. This allows changes in plan for one state to influence the plan for others, guaranteeing convergence toward a coherent, reward-maximizing policy (see Sutton & Barto, 1998).

Something interesting happens when this kind of iterative procedure is introduced within the probabilistic inference framework we have taken as our focus. Starting with the reward query, as we have defined it, the natural approach is to perform this form of query repeatedly, using the
posterior distribution $p(\pi|s,u_0)$ obtained on each iteration $n$ as the prior distribution $p(\pi)$ for $n+1$, applying Bayes’ law to compute:

$$p_{n+1}(\pi|s,u) \propto p(u|s,\pi)p_n(\pi|s,u).$$

(5)

In the Appendix, we provide formal proofs showing that this approach gives rise to optimal policies. From a psychological and neuroscientific standpoint, however, the approach also has another source of appeal: It turns out to be formally identical to currently prevalent behavioral and neural models of perceptual decision making. Most such models center on sequential sampling, an iterative procedure in which decision parameters are repeatedly updated based both on perceptual observations and on the interim values of the same parameters from the previous iteration (Beck & Pouget, 2007; Bogacz et al., 2006; Gold & Shadlen, 2007; Rao, 2006). The recursive procedure captured in Equation 5 is precisely analogous. The role of the perceptual observations is simply taken by the ‘observation’ $u_0 = 1$.

When implemented in this computationally adequate iterative fashion, the reward query appears to provide a powerful and general method for goal-directed decision making. Unlike the state query, it takes reward information directly into account, and retains a sensitivity to alternative outcomes. Unlike the policy query, it avoids the need to search, in a serial fashion, through candidate policies, relying on ‘bottom-up’ selection of policies rather than their top-down imposition. And again unlike the policy query, the reward query squares with the intuitive idea that goal-directed decision making works as much in the backward direction, from outcomes to actions, as through forward search.

Summary

We have outlined a set of inference procedures for goal-directed decision making, each applying a particular kind of query to a causal model of action. As stated at the outset, we do not identify goal-directed decision making with any single one of these procedures. Instead, we suggest that goal-directed decision making rests upon a suite of procedures, any subset of which may be involved in a particular decision. Indeed, the set of possible procedures is not limited to the ones we have listed above. For example, under the general framework it would be possible to select policies by evaluating $p(s'|s,\pi)$, aiming to maximize the probability of a particular outcome state; or to select a policy by conditioning on both state and reward variables (see Attias, 2003; Verma & Rao, 2006b).

While portraying goal-directed decision making in this multi-faceted way, we have also identified strengths and weaknesses among alternative procedures, calling particular attention to a bottom-up, reward-driven approach. This reward-query based form of decision making will provide a focus in much of the remainder of the paper, partially given its relative novelty and partially given its fit with empirical findings to be reported below.

In characterizing various forms of query, we have written down several closed-form mathematical expressions. However, it is important to note that the kinds of inference we have discussed are readily implemented using local, distributed algorithms. Recent work has shown how at least one such algorithm, the belief propagation procedure (Pearl, 1988) and its close relatives, could be implemented in neural hardware (Beck & Pouget, 2007; Koechlin, Anton, & Burnod, 1999; Lee & Mumford, 2003; Litvak & Ullman, 2009; Rao, 2006), an important point in evaluating the biological plausibility of the present framework.
Simulation Studies

Having outlined the general theory, we now report a series of simulation studies that illustrate how it would account for specific hallmark examples of purposive behavior. The first set of simulations to be reported focuses on the selection of single actions, and models benchmark phenomena in this context including incentive devaluation effects. These simulations also illustrate the ability of the framework to cope with probabilistic outcomes. A second set of simulations addresses the case of sequential decision, addressing how cost-benefit analyses are performed in this context and applying the computational framework to detour behavior and latent learning. As signaled earlier, the main focus in the present simulations, as well as the experiments that follow, is on the reward-query mechanism.

General Methods

Representations and Architecture

All simulations were run using the Matlab Bayes Net Toolbox (Murphy, 2001), combined with custom Matlab (Mathworks, Natick, MA) code, which is available upon request.

Simulations addressing single-step decisions employed the architecture from Figure 2C. Multi-step tasks were modeled using the architecture from Figure 2D, extended to include the minimum number of actions required for the task simulated. As explained in earlier sections, states, actions and policies were represented by discrete, multinomial variables. Policies, in particular, were modeled using a set of nodes connected to each action variable, with each node representing the policy for a single state. Each policy-node value corresponded to a unique, deterministic policy for the relevant state. As described earlier, reward was modeled using a binary variable connected to each state variable. For each task modeled, a scalar reward value \( R(s) \) was assigned to each state \( s \). The resulting set of reward values was then scaled to fall between zero and one and used to define the CPD for the reward variable, using the following linear transformation:

\[
p(u|s) = \frac{1}{2} \left( \frac{R(s)}{r_{\text{max}}} + 1 \right), \quad r_{\text{max}} = \max_j \{R(s_j)\}. \tag{6}
\]

In simulations involving action sequences, a node \( u_G \) was included, as shown in Figures 2 and 4:

\[
p(u_G) = \frac{1}{N} \sum_i p(u_i) \tag{7}
\]

where \( N \) is the number of \( u \) nodes. Note that the resulting CPD does not discount rewards based on their latency. Nor are negative rewards weighted more heavily than positive ones. However, temporal discounting and loss aversion could be incorporated into the model with minor changes to the expression above.

Simulation Procedure

Each simulation involved imposing a set of values on one variable or set of variables and computing the posterior distribution over another variable or variables. In all cases, posterior
probabilities were computed using the junction tree algorithm (see Jensen, 2001). Where state or reward queries were used, the iterative procedure described above, and detailed in the Appendix, was followed.

**Simulation 1: Instrumental Choice**

We begin with a set of simulations involving selection of single actions. Simulation 1A models a simple instrumental choice situation, based on experiments with rats. Simulation 1B simulates a choice task involving probabilistic outcomes, based on experiments with humans.

**Simulation 1A: Choice, Devaluation and Contingency Degradation**

We model here a scenario like the one used in the Introduction to illustrate devaluation, and used in Balleine and Dickinson (1998d), along with numerous other studies cited earlier. Here, a rat has access to two levers, positioned to the left and right of its cage. Pressing the left lever yields one kind of food, and pressing the right another. We assume that, at baseline, the rat prefers the food associated with the left lever, assigning a reward value of 2 to this food and a value of 1 to the other. The situation is modeled by defining three states, no food (the initial state, reward value zero), food 1 and food 2; and two policies, press left and press right, matched with corresponding actions.

**State, policy and reward queries.** Our first set of simulations addresses the simple question of how the rat decides, based on its knowledge of the causal structure of the environment and its preferences over outcomes, which lever to press. As indicated in earlier discussion, we suggest that there may be multiple routes to this decision, involving different forms of model-based inference. If we are willing to imagine that the rat begins with the goal of obtaining its preferred food (as sometimes posited to account for the phenomenon of Pavlovian-instrumental transfer, for example; see Holland, 2004), then the choice can be modeled based on a state query: The variable representing the outcome state (s’ in Figure 4, top left) is treated as observed, and the posteriors over policies are computed. Because the preferred food is dispensed if and only if the left lever is pressed, the posterior probability of the press left and press right policies, as computed by the model, turn out to be 1 and 0, respectively. The same decision can also be modeled as based on a policy query. Here, the policy is treated as observed, first set to press left, then separately to press right. In both cases, the inference procedure yields specific posterior probabilities at the reward node. The probability \( p(u|x) \) turns out to be larger under the press left policy (1.00) than under press right (0.75), providing a sufficient basis for choice.

Finally, the decision can also be made based on a reward query. Here, the reward variable is treated as observed \((u = 1)\), and inference yields posterior probabilities for the two available policies. Figure 5A shows the evolution of these posteriors, over iterations of the inference algorithm. Also shown is the expected value of the current mixture of policies (the average of \( p(u|x) \), weighted by the posterior probability of \( x \) on the current iteration). As the figure shows, over time, the model converges on the optimal deterministic policy.

**FIGURE 5 AROUND HERE**

**Devaluation.** In Balleine and Dickinson (1998c), as in numerous related studies, the incentive value of one action outcome was devalued by specific satiety, leading to an immediate reduction in performance of the associated action. This devaluation effect can be captured in our model by simply changing the reward value associated with one food outcome. To simulate the effect of this, we reduced the reward value associated with the formerly preferred food from
two to zero. Note that this change directly affects only the CPD of the reward variable; \( p(u|s') \) is reduced for the state of \( s' \) representing the devalued food — consistent with the reduction in OFC activation observed following incentive devaluation (see, e.g., Valentín et al., 2007). When a reward query is performed, however, the impact of this local change propagates to the level of the policy node, yielding a reversal in choice (Figure 5B).

**Contingency degradation.** A key finding from the animal learning literature involves the consequences of breaking the causal contingency between a specific action and outcome. Here, typically, the animal first learns to associate delivery of a certain food with a particular action, but later begins to receive the food independent of the action. The impact of this change is that the animal less frequently produces the action in question (Colwill & Rescorla, 1986; Dickinson & Mulatero, 1989; Williams, 1989). This effect was simulated in the lever choice context as before, with the original reward assignments (2 versus 1). To model contingency degradation, the CPD for \( s' \) was modified so that pressing the left lever resulted in the no food state. Choice was then modeled through a reward query. In this context, the local change at the level of \( s' \) again propagates to the policy nodes, yielding press-right (Figure 5C).

**Simulation 1B: Probabilistic Outcomes**

As discussed earlier, it is often important in decision making contexts to weigh a range of alternative action outcomes. Because the present theoretical framework is framed in terms of probabilistic inference, it extends naturally to such situations when appropriate queries are used. In order to model such a situation, we employed a simplified version of the Iowa Gambling Task, a laboratory task used in many studies of human incentive choice (Bechara, Damasio, Damasio, & Lee, 1999; Bechara, Damasio, Tranel, & Damasio, 1997; Bechara, Tranel, Damasio, & Damasio, 1996). Here, we consider a subject choosing between two decks of cards. Each card indicates a monetary gain or loss, and the subject’s task is to maximize cumulative gains. In one deck (left), half of the cards confer a gain of 1 and the other half no gain or loss. In the other deck (right), half the cards confer a gain of $2 but the other half a loss of $3. The optimal choice is obviously the left deck, which has a lower maximum but larger expected reward.

The task situation was modeled using five states, **pre-choice**, +$1 card, $0 card, +$2 card and -$3 card; and two actions, **left deck** and **right deck**. Reward values were assigned to the states in the obvious way, and the appropriate (stochastic) transition function was implemented. Posing a reward query to the resulting model yielded the decision trajectory shown in Figure 5D, converging on the optimal policy.

**Simulation 2: Sequential Decision**

**Simulation 2A: Two step T-maze**

As an example sequential-choice situation, we focus in this simulation on a two-step T-maze scenario described by Niv, Joel and Dayan (2006), discussed briefly in the Introduction. Here, as shown in Figure 6A, a rat must make a sequence of two left-right decisions, arriving by these at a location containing either cheese, carrots, water, or nothing. The states included in our model of this situation include the four outcome states as well as the three preceding choice points (see figure). We assume (following Niv et al., 2006), that the animal prefers cheese over carrots and carrots over water, implementing these assumptions by imposing the reward values cheese 4, carrots 3, and water 2 (with the empty maze arm assigned reward zero).
Sequential decision. Figure 6B shows the decision trajectory under a reward query. The model converges on the sequence left, left, a policy that takes it to the preferred cheese reward.

One detail of the plot in Figure 6B is worth remarking for later reference. Note the interesting trajectory of the solid blue and green traces, which relate to the decision at S1. Although the decision ultimately tips toward left, early on there is transient movement toward right. The explanation for this lies in the recursive structure of the planning problem. As discussed earlier, the optimal decision at an early step in a plan can depend on what is planned for later steps. This is true for the rat in the two-step T-maze. At the outset of planning, the rat has no established plan for what to do at S2 or S3; he is equally likely to go left or right at those points. As a result, the expected reward for a left turn at S1 is (4+0)/2, and for a right turn it is (2+3)/2. Since the latter quantity is larger, the (locally) optimal choice at S1 is to go right. Eventually, as better plans emerge for S2 and S3, the expected reward for left and right turns at S1 move toward 4 and 3, respectively, making it better to go left at S1. The resulting decision-making reversal, readily legible in Figure 6B, translates into an empirically testable prediction of the model, to which will return in subsequent discussion.

Motivational state. A common manipulation in empirical studies of goal-directed behavior involves inducing a shift in motivational state (Balleine, 1992; Balleine & Dickinson, 1994; Dickinson & Dawson, 1989). Niv and colleagues (2006) used their two-step T-maze scenario to discuss the impact of such shifts on incentive value. They considered a situation in which fluid deprivation is used to make the rat thirsty. This, they reasonably suggested, would induce a change in the reward value associated with the four outcomes. In order to model this situation, we changed the reward values in our model to the ones posited by Niv and colleagues (2006): cheese 2, carrots 1, and water 4. These changes directly affect only the two reward nodes (u), changing p(u|s) in a fashion intended to be analogous to the shifts in BLA and OFC responses with changes in motivational state (De Araujo, Kringelbach, Rolls, & McGlone, 2003; LaBar et al., 2001). Applying a reward query yielded an entirely new action sequence, as show in Figure 6C.

Effort-based decision. A key feature of reward-based decision making is the ability to weigh the rewards associated with outcomes with the expected costs of obtaining them. A clear example of this kind of cost-benefit analysis comes from work on effort-based decision making. A number of studies have examined their behavior when placed inside a T-maze where both arms contain food, but one also contains a scalable barrier that the animal must surmount to access the food reward. The common finding is that, unless the reward on the barrier side is larger by a sufficient degree, animals will forgo it, avoiding the effort required (Salamone, Correa, Farrar, & Mingote, 2007; Walton, Kennerley, Bannerman, Phillips, & Rushworth, 2006).

This sort of cost-benefit analysis can be modeled very naturally within the present framework. For simplicity, we do so using the two-step T-maze scenario already established. Here, we re-impose the original reward values on the outcome states, but also imagine that there is now a scaleable barrier placed at S2 (see Figure 6A). The cost of scaling this barrier is inserted into the model by simply changing the reward value associated with S2 to -2. Applying a reward query to the model under these circumstances yields the decision trajectory in Figure 6D, which reflects the inference that the value of the most preferred reward is not worth the effort cost required to obtain it.
Simulation 2B: Ostlund, Winterbauer and Balleine (2009)

The two-step T-maze described by Niv and colleagues (2006), our focus of Simulation 2A, provides a convenient setting for considering several aspects of sequential decision. However, it is not a setup that has been employed in actual empirical studies. In this simulation, we turn to a recent experiment, which shares the same underlying structure. Ostlund, Winterbauer and Balleine (2009) trained rats to execute two-step sequences in order to obtain food rewards. The rats had access to two levers. When a rat pressed the right lever and then the left, a bit of sucrose was delivered. When the levers were pressed in the opposite order, the rat received polycose (of course, for other rats the assignment of foods to sequences was reversed). Following training, one of the foods was devalued through satiety. When presented with the two levers in this setting, rats tended to execute the sequence yielding the non-devalued food more frequently than the opposite sequence.

Although this experiment involves lever-pressing rather than maze navigation, the form of the decision problem aligns precisely with the two-step T-maze from Simulation 2A (see Figure 6A). State $S_1$ now corresponds to the rat’s initial situation, facing the two levers, with available actions press-left and press-right. State $S_2$ corresponds to the rat’s situation after having pressed the left lever once; state $S_3$ the situation after pressing right once. The outcomes for press-left and press-right are, respectively, nothing (reward 0) and polycose (reward 1) at $S_2$, and sucrose (1) and nothing (0) at $S_3$. Using the model from Simulation 2A to simulate these initial conditions leads to selection of the sequences left-right and right-left (with equal probability) over left-left and right-right (Figure 7). Simulating devaluation by reducing the reward associated with sucrose to 0.5 leads to a preference for left-right over all other sequences, in line with the empirical observation (Figure 7).

FIGURE 7 AROUND HERE

Ostlund and colleagues (2009) also showed analogous changes in sequence production following contingency degradation. Simulating contingency degradation in the present model, using the approach adopted in simulation study 1A yields parallel results (data not shown).

It is worth remarking that the computational account we are offering here for the findings of Ostlund et al. (2009) differs from those authors’ own interpretation. Ostlund and colleagues considered the observed pattern of behavior to indicate the involvement of “chunked” representations of action sequences. The present simulation illustrates that chunking is not in fact necessary. Having noted this, however, we hasten to add that ‘chunked’ representations are nonetheless likely to play a role in goal-directed decision making, a point to which we shall return in the General Discussion.

Simulation 2C: Detour Behavior

To model detour behavior, we divided the plan of the maze from Tolman and Honzik (1930) into a set of nine discrete zones, making these the states in our model (Figure 8A), and included three policies (with corresponding actions): straight, left, and right. A reward of 1.0 was linked with the final state ($S_5$), and in order to encourage shortest-path behavior, a small negative reward (-0.1) was associated with all other states. Applying a reward query yielded the decision trajectories at $S_2$ shown in Figure 8B. Placement of a barrier at location $a$ in Figure 8A was modeled by changing the CPD for the state variables, such that execution of the straight action at $S_2$ yielded $S_2$ again as the outcome. With this local change, a reward query yielded the
decision to follow the upper passageway in the figure (Figure 8C). Placement of the barrier at location \( b \) (see Figure 8A), modeled in the same manner, yielded the decision to follow the lower passageway, as in Tolman and Honzik (1930).

**FIGURE 8 AROUND HERE**

**Simulation 2D: Latent Learning**

In order to model latent learning, we divided the plan of the maze from Blodgett’s (1929) study into discrete locations as shown in Figure 9, using these as the states in our model. Two actions were included, *left* and *right*. At blind-alley locations (e.g. \( S_2 \)), either action resulted in no movement (i.e., a return to the same location on the next time-step). As in the previous simulation, a small negative reward (-0.1) was associated with all states. A reward query under these circumstances yielded random choices at each T-junction. In order to model the introduction of the food reward, the reward value associated with \( S_{13} \) (see Figure 9) was increased to 1.0. This local change yielded completely different model behavior. A reward query yielded the set of T-junction decisions shown in the figure, simulating the post-reward behavior reported by Blodgett (1929).

**FIGURE 9 AROUND HERE**

**Simulation Studies: Discussion**

Having put forth a framework for goal-directed decision making, we have shown in the foregoing simulations how it applies in a number of canonical behavioral settings, including simple incentive choice, risky choice, multi-step planning, devaluation, contingency degradation, change in motivational state, detour-finding and latent learning. While our models align broadly with existing, informal interpretations of these phenomena, they go beyond those interpretations, offering a formal account within which the integration of incentive and contingency knowledge is explicitly modeled, and providing a computational framework for understanding how specific brain regions may collaborate to generate goal-directed decisions.

Our simulations were intended to provide a proof of concept; they cannot of course substitute for empirical tests of the theory. Fortunately, the framework we have presented gives rise to a number of testable predictions. In the General Discussion, we will enumerate a set of neuroscientific predictions, potentially testable through single-unit recording methods. At the present juncture, we focus on two behavioral predictions, experimental tests of which are reported in the next section.

The first of these predictions relates to the decision-making reversal we noted in discussing the results of Simulation 2A, illustrated in Figure 6B. This reversal took place at \( S_1 \), the first decision point in the two-step T maze, where the decision tipped slightly toward *right* before ultimately converging on *left*. As discussed earlier, this reversal reflects the recursive structure of multi-step planning, where the proper decision concerning early actions depends on what later actions are planned. In the simulation, early in the decision-making process, no decision has yet been made about what to do at \( S_2 \) or \( S_3 \). Given this setting, the optimal choice is to go right at \( S_1 \), since the mean utility of the outcomes in that direction is higher than the mean utility in the other direction. Once the policies at \( S_2 \) and \( S_3 \) have matured, the decision at \( S_1 \) goes left, reflecting the fact that the maximum outcome utility in that direction is higher than the maximum on the right. With these points in mind, we refer to the conditions that give rise to the reversal in
Simulation 2A as *mean-max conflict*. Below, in Experiment 1, we tested the prediction that mean-max conflict should give rise to decision reversals in human decision making.

A second prediction stems from an aspect of our models that has not yet been broached. Recall that with each iteration of the decision-making process, for each planned action, our models compute a posterior probability distribution over outcomes. It turns out that, in the setting of a reward query, this posterior distribution is *optimistic*. That is, it is weighted toward high-utility outcomes. An example is provided by the model presented in Simulation 1C. Here, when the model decides to pick from the left deck (as in Figure 5D), the objective probability of a monetary gain is 0.50. However, at asymptote, the model attaches to this outcome a posterior probability of 0.57. This effect stems from the nature of the reward query: Decision-making in this setting is driven by the *assumption* that $u = 1$. This assumption feeds into the calculation of outcome probabilities, with the natural consequence that they are weighted toward states with higher utility. It is important to note that this aspect of the model does not affect the model’s actual decisions. That is, the model’s choices of action conform to sound calculations of expected utility. Nevertheless, even as the model chooses rationally, it gives rise to optimistic estimates of outcome probability. This provides a second key behavioral prediction, which we test in Experiment 2 below.

**Experimental Studies**

*Experiment 1: Reversals in Sequential Decision*

The previous section described a set of task circumstances that, in our model, lead to an interesting kind of decision reversal, a situation we labeled mean-max conflict. In the present experiment, we tested for the same dynamic in human decision making.

Participants performed a two-step decision making task, designed to engender mean-max conflict. Throughout the experiment, the participant dealt with a set of five abstract symbols, each of which was associated with a different point value. On each trial, a subset of four of these characters was presented, and the objective was to earn as many points as possible by selecting one character from among the set displayed. Critically, selection occurred in two stages. The initial choice display featured two boxes, each containing two characters (Figure 10A). By means of a keypress, participants selected one of the two boxes. With their second response, participants selected between the two characters within the remaining box.

**FIGURE 10 AROUND HERE**

The experiment included two trial-types. On *match* trials, the box with the larger mean character value also contained the single character with the highest point value. For example, one box might contain characters worth nine and three points, the other box characters worth eight and zero points. *Mismatch* stimuli placed the most valuable character in the box with the smaller mean. For example, one box might contain characters worth nine and zero points, and the other box ones worth eight and three points. Because mismatch imposed a mean-max conflict, as we have defined it, we predicted that these trials should induce a pattern of response reversal like that induced by mean-max conflict in our simulations.

For illustration, we simulated both a match and a mismatch trial, using a reparameterization of the model from Simulation 2A. State 1 was understood as the initial display, with actions *left* and *right* leading probabilistically — as in the experiment — to state 2 (left box remaining) and state
3 (right box remaining), each with available actions top and bottom. Figure 10C plots the evolving posterior probability of the correct box-choice action. Note the early dip in the value of the posterior on the mismatch trial.

The plot bears a meaningful resemblance to one from a study by Gratton and colleagues (Gratton, Coles, Sirevaag, Eriksen, & Donchin, 1988), reproduced in Figure 10B. These investigators predicted a very similar response reversal, albeit with very different origins and in a very different task setting (the flanker paradigm introduced by Eriksen & Eriksen, 1974). In order to test for the predicted reversal, Gratton and colleagues (1988) imposed pressure on participants to respond quickly, sorted responses into reaction-time bins, and calculated mean response accuracy in each bin. Their correct prediction was that in the condition involving response reversal, response accuracy at low response times would fall below chance.

We applied this same approach in our experiment, examining response accuracy over a range of reaction times. Based on the behavior of the model, we predicted that on mismatch trials, at low reaction times, the proportion of correct responses would fall below the chance level of 0.5.

Method

Participants. Fourteen neurologically normal individuals from the Princeton University community (18-40 years) participated, receiving a nominal payment in return. All participants provided informed consent. All procedures were approved by the Princeton University Psychology Department Institutional Review Board.

Task and Materials. The task was programmed using the Psychophysics Toolbox (Brainard, 1997) for Matlab. The five symbols included in the stimuli in the experiment were kanji pseudo-characters, which included the four shown in Figure 10A. Characters were associated with point values of 0, 3, 8, and 9, with the character-value pairing varying across participants.

At the start of each trial, four characters were displayed, positioned within two horizontally arrayed boxes, again as in the figure. The participant chose a box by pressing a spatially aligned key (s or d), using the left index or middle finger. On 75% of the trials this caused the unchosen box to immediately disappear. On the remaining trials, the chosen box disappeared. Next, the participant chose between the two characters in the remaining box, again using spatially aligned keys (b and d) and the right index and middle fingers. The point value for the chosen character was then displayed, and added to a running total of points earned, which appeared at the periphery of the display. A 750 msec inter-trial interval then followed.

To impose time pressure on the initial response, a deadline was enforced. One of three cues appeared prior to each block of trials (see below for details of block structure), either the word “shorter,” “medium,” or “longer,” indicating deadlines of 500, 1000, and 1500 msec respectively. If no response was completed before the prevailing deadline, the boxes and characters disappeared, a brief tone sounded, and a 1500 msec inter-trial interval followed.

The characters appearing on each trial, and their distribution between boxes, were chosen so that one half of all trials featured match stimuli (as defined above) and half mismatch stimuli. Character sets were also crafted to assure that characters with values of nine and ten occurred equally often as targets. Details of character-set frequency are presented in Table 1.

TABLE 1 AROUND HERE
**Procedure.** At the beginning of the session, participants were shown each of the five characters and informed of its associated point value. Participants then performed 16 practice trials, identical to those already described but with no response deadline. The entire training procedure was repeated until accuracy (proportion of trials where the participant made point-maximizing responses) surpassed 80%. The participant was then oriented to the deadline cues. The experiment proper included 504 trials divided into blocks of 24, each led by a deadline cue. The order of deadline conditions was such that no condition repeated before all three had occurred, and no two adjacent blocks carried the same deadline. Within each block, stimuli were drawn randomly (with replacement) based on the probabilities in Table 1.

**Analysis.** Analysis focused on the accuracy of responses completed prior to deadline. A two-tailed paired t-test was used to compare response accuracies between match and mismatch trials. The central prediction, however, was that mean accuracy for mismatch trials, but not match trials, would fall below chance at low RTs. To test this, response data were sorted according to stimulus type (match vs. mismatch) and by reaction time (RT), using 100 msec bins, pooling across deadline conditions. Bins where at least one participant contributed no data were excluded from further analysis. The mean accuracy of the first response (box choice) was calculated for each stimulus type within each RT bin. Using two-tailed single-sample t-tests, response accuracies were compared to chance (0.5). Corrections for multiple comparisons took into account the specific prediction that below-chance responding would occur in a low RT bin. Tests began at the lowest RT bin, and proceeded in order of ascending RT. No correction was made on the first test. On the second test, Bonferroni correction was made by multiplying the \( p \) value by two, on the third test by three, and so forth.

**Results.** Overall response accuracy was higher for match trials (0.69) than mismatch trials (0.66), \( t(13) = 2.27, p = 0.04 \). More importantly, consistent with our central prediction, accuracy for mismatch trials was found to be significantly below chance in the lowest RT bin, covering 201-300 msec \( (t(13) = -2.77, \text{two-tailed } p < 0.05) \). Figure 10D shows mean accuracy in this RT bin, as well as in the adjacent time bin and the highest RT bin (accuracies in the remaining bins were intermediate between those shown).

**Discussion**

The present experiment confirmed a specific prediction of the proposed computational paradigm. The prediction stems from the recursive structure of multi-step plans, according to which the optimal choice at early steps depends on what's planned for later steps. In our framework, decision making at early steps does not wait for the plan for later steps to be settled, and as a result reversals can occur in the selection of early actions. We focused on a specific setting in which this is predicted to occur (mean-max conflict), and observed choice behavior consistent with the prediction.

It should be noted that the pivotal finding in this experiment — below-chance accuracy on mismatch trials — cannot be explained by a sampling scenario in which participants encoded some subset of the four items presented and chose the box with the best of the characters in that subset. As examination of Table 1 will confirm, such a procedure might result in higher accuracy on match than mismatch trials, but it would not yield below-chance responding.

In evaluating the implications of the present experiment, it is of course important to consider whether the same results would be predicted by other computational approaches to goal-directed decision making. This issue will be addressed later, in conjunction with a wider discussion of alternative models. For the moment, we note that the results obtained in Experiment 1 appear inconsistent with serial, subgoaling-based models (e.g., production system models), where a goal is first identified, and only then are actions selected to obtain that goal.
The results of Experiment 1 point instead to the idea that decision-making at multiple stages of a multi-step plan can occur in parallel.

**Experiment 2: Optimistic State Prediction**

This experiment tested the second prediction drawn from our simulation studies, namely that goal-directed decision making will lead to an overestimation of the probability of desirable outcomes. To test for this optimism effect, we had participants choose between visual stimuli each with a different probability of yielding a (constant-valued) ‘win’ outcome. In choosing, participants used a visual analogue scale (VAS) to indicate their ‘gut feeling’ concerning the likelihood of a win outcome, given the character being chosen (see Figure 11A). To provide a baseline for comparison, on some trials participants were only shown one character, which they were obliged to ‘select,’ again by using the VAS. We assumed that the choice condition (two characters) would elicit goal-directed decision making, whereas the forced condition (one character) would not. Based on this, we predicted that VAS ratings would be higher in the choice condition than the forced condition.

**FIGURE 11 AROUND HERE**

**Methods**

*Participants.* Participants included 28 neurologically normal individuals from the Princeton University community (22 female, median age 19). All participants were fluent in English, and did not have Japanese language knowledge. Eleven participants were paid on an hourly basis at a rate of $12 per hour, and the others were granted participation credit as compensation. In addition, all participants received a bonus payment of eight dollars. Participants provided informed consent prior to the beginning of the experiment. All procedures were reviewed by the Princeton University Institutional Review Panel.

*Task and Materials.* Stimuli included five black Japanese Hiragana characters on a plain white background, displayed as images measuring approximately 400 pixels by 400 pixels on a personal computer. Each character was mapped to one of the following five reward probabilities: 0.20, 0.35, 0.50, 0.65, and 0.80. The mapping between stimuli and reward probabilities was randomly assigned for each participant.

Characters were integrated into two classes of display, defining two trial-types: *choice* trials and *forced* trials. On choice trials, participants were presented with two characters. Next to each character, a vertically oriented rectangle was displayed (Figure 11A). Participants selected one of the two characters by using a mouse to click inside the adjacent rectangle, using it as a VAS. When the cursor was placed over the inside of the bar, it was filled with a blue block from the bottom to the height of the point of the cursor. The height of this bar followed the cursor until the mouse button was clicked (at which point feedback was displayed), or the cursor was no longer over the bar (at which point it was displayed as an empty frame again). Participants were instructed to indicate their “gut feeling” about whether or not a win outcome would result on that trial, using the VAS bar. Participants were told to click at the bottom of the bar if they felt certain they would not receive a point, in the middle if their gut feeling was completely ambiguous, and at the top of the bar if they felt certain they would receive a point. Forced trials were the same, except that only one character and one (adjacent) slider bar were presented. Participants were obliged to select the character presented, again indicating their “gut feeling” about the probability of a win outcome.
Immediately following the participant’s VAS response, the computer displayed feedback after the participant made his or her response. If a character paid a ‘point’ on that trial, a dollar sign (‘$’) was displayed below that character. If a character did not pay out, an ‘X’ was displayed below the character. This feedback was in color (green for $, red for X) for the character to which the participant responded, but also in black for characters that were not chosen. Points won were added to a running total, which appeared at the top of the display. To continue on to the next trial, the participant was prompted to click on a small grey button in the center of the screen.

Procedure. The experiment consisted of one session approximately one hour in length. Participants first received verbal instructions and performed two sample trials: one choice, one non-choice. The characters used in these sample trials were other Hiragana characters not used in the experimental trials. Participants were informed that every character they would encounter during the experiment had its own fixed probability of a payout of one point, and that the goal was to earn as many points over the course of the experiment as possible. Participants were informed they would receive a bonus payment based on their score, but were not told how much money each point would be worth.

The experiment proper comprised 300 self-paced trials, presented continuously. Stimulus order was determined by forming a series of blocks of thirty pairs of characters, a random ten of which were designated forced trials. The twenty choice trials included a random ordering of all pairings of item and position, with the restrictions that a single item could never appear twice in the same display and that no block could start with the same pair of characters presented at the end of the preceding block. Non-choice trials were based on the set of all possible character pairings. For each pairing, only one of the two relevant characters was displayed, with display laterality chosen randomly. The character was chosen to match the one selected by the participant on the most recent choice trial in which the same pair of characters was presented. In cases where the pair had not yet been presented, the display character was selected at random. This procedure was adopted in order to match the frequency of each individual character on forced trial to the same character’s selection on choice trials.

Analysis. The first 100 trials in each session were treated as a learning period, and were not included in data analysis. For all trials, the character chosen by (or for) the participant was recorded, as was the location of that character (left or right). Participants’ ratings concerning of the probability of payout were recorded as a proportion of the VAS that was ‘filled.’ These ratings were subjected to a two-way within subject ANOVA with factors for item (character) and choice condition. This was followed up by five paired t-tests, each comparing mean ratings in choice versus forced conditions for a specific character. Results of these tests were evaluated at a significance threshold of \( p = 0.05 \), two-tailed. No correction for multiple comparisons was made, given the existence of \textit{a priori} predictions in all cases.

Further analyses were performed to rule out an alternative explanation for the predicted effect. It might have been the case that participants’ estimates of character-specific payoff probabilities varied over some range — either randomly or based on the recent sampling history — with the distributions for these ratings overlapping highly between characters. If this were the case, and if participants followed a policy of choosing the character with the highest estimated payoff probability on choice trials, the result would be higher ratings on choice than forced trials, simply as a consequence of sampling bias. Note that the impact of this sampling bias on the predicted optimism effect would depend entirely on the degree of overlap among the distributions for individual characters. A high degree of overlap would presumably be detectable in a participant’s choice behavior, manifesting as a high degree of choice variability. Thus, the
sampling bias explanation predicts a positive correlation between choice variability and optimism. We tested for this correlation, quantifying variability in terms of choice entropy \(-p \log(p)\), where \(p\) is the probability of choosing the more valuable character on choice trials. The optimism effect was quantified as the mean difference in VAS rating, across characters, between choice and forced conditions. A second test drew on the observation that, under a sampling bias explanation, the mean VAS rating for any particular character on choice trials should depend on the identity of the rejected character; higher ratings should occur when the rejected character is closer to the chosen character in payoff probability. In order to evaluate this, we focused on trials where participants were offered, and chose, the most valuable character (payoff probability 0.80). Using a one-way repeated measures ANOVA, the mean VAS rating for this character was compared across cases where it was paired, respectively, with each of the other characters in the experiment.

**Results**

As intended, the frequency with which each character appeared on forced trials was closely matched to the frequency with which that character was chosen by participants on choice trials (Figure 11B). Mean accuracy on choice trials (proportion of trials where the more valuable character was selected) was 0.78.

Ratings (VAS responses) for all five characters, in both forced and choice conditions are shown in Figure 11C. A two-way ANOVA, based on the 24 participants who sampled all five characters, yielded a main effect of character \((F(4,20) = 8.36, p < 0.001)\), a main effect of condition \((F(1,23) = 14.22, p < 0.002)\). The interaction between these factors did not reach statistical significance \((F(4,20) = 2.17, p < 0.08)\). Character-specific t-tests comparing ratings between forced and choice conditions confirmed significant differences in all five cases \((t(23) > 2.10, \text{two-tailed } p < 0.05)\).

As noted under Methods, we conducted two analyses to evaluate whether the effect of choice condition on ratings might have reflected a sampling bias, stemming from variability in character-specific outcome probability estimates. Neither analysis supported this alternative explanation. No significant correlation was observed between choice entropy and optimism, as defined under Methods \((r = 0.09, p = 0.67)\). Furthermore, on choice trials where participants were offered the most valuable character (payoff probability 0.80) and selected this character, mean VAS ratings did not vary appreciably with the identity of the rejected character (Figure 11D). A one-way ANOVA showed no main effect of rejected character \((F(3,24) = 0.6, p = 0.62)\).

**Discussion**

This experiment confirmed a second specific prediction of our computational paradigm, which is that goal-directed decision making should give rise to optimistic predictions concerning the likelihood of desirable outcomes. As in the case of Experiment 1, it is important to consider whether the result obtained might have been predicted by other algorithms for goal-directed decision making. In the case of the present experiment, the answer appears to be no, but we postpone an explanation of this point for the General Discussion.

The optimism effect predicted and confirmed in the present experiment bears an interesting relationship to what previous work has labeled the “illusion of control.” Here, individuals make more optimistic outcome predictions when their actions are freely chosen than when they are dictated (Presson & Benassi, 1996). For example, Langer (1975) found that experimental participants expressed greater confidence in their chances of winning a drawing when they were
permitted to select a ticket from among a set of objectively equivalent tickets than when a random ticket was simply given to them. A standard explanation for this effect has been that choice serves as a cue falsely implying outcome controllability (Langer, 1975; Presson & Benassi, 1996). The present work suggests a different, though perhaps not incompatible, explanation, which is that choice gives rise to optimism as a natural consequence of the computations involved in goal-directed decision making. Further evaluating the contribution of this mechanism to the illusion of control presents a target for additional experimentation.

General Discussion

The distinction between goal-directed and habitual behavior is time-honored and empirically well established in both psychology and neuroscience. Since the days in which purposivism was pitted against behaviorism, tremendous progress has been made toward understanding both modes of decision making. However, there is also a striking imbalance in progress toward specifying computationally explicit theories in the two domains. Whereas research on habitual, stimulus-response control has begun to orient around a specific set of computational formalisms, research on goal-directed decision making has not yet identified an analogous set of heuristic computational tools.

In the present paper, we have advanced a computational account of goal-directed decision making. The theory takes, as its point of departure, the idea that the nature of causal knowledge, in both humans and higher animals, can be understood in terms of probabilistic graphical models and the procedures they support. This idea is brought together with a recently emerging machine learning approach to the task of solving Markov decision problems, which treats it as a problem of probabilistic inference.

The resulting account asserts, most fundamentally, that goal-directed decision making relies upon a rich generative model of action, which links actions to outcomes and states to reward values. We identify each sector of this model with a specific, though diverse, set of cortical and subcortical structures comprising, respectively, a policy system, state and action systems, a state-projection system, and a reward system. We have characterized goal-directed decision making as a family of operations performed on this generative model, each involving a different specific kind of query, but all working through probabilistic inference to arrive at a policy for action. Among the several forms of query delineated, one that appears to be of particular novelty and interest is what we have termed a reward query, in which optimal policies are inferred based on assumptions concerning reward.

We have presented simulations illustrating how the theory would account for a range of benchmark behavioral phenomena, and reported experiments in which specific predictions of the framework were tested and confirmed. In conclusion, we briefly consider the relationship between the ideas we have presented and previous work, and enumerate some areas for further development.

Relation to Psychological Accounts

As discussed at the outset, the present work ties into a long tradition beginning with Tolman and other early purposivists, within which goal-directed decision making is understood as involving the integration of causal knowledge and incentive knowledge to infer rewarding policies for action. Our contribution is to provide a computational account of how each form of knowledge is represented, and of how their integration is accomplished. The specifics of this account link it to a wide range of research in psychology using probabilistic graphical models as a medium for
explaining human and animal decision making. The closest connection, as noted earlier, is with recent work on causal knowledge, but the same conceptual tools have been applied in work in areas ranging from perception to social cognition.

As discussed further below, the account we have presented is normative, in the sense that it proposes that human and animal decision making results in actions that are optimal under some criterion, most centrally the criterion of reward maximization. This characteristic ties the present theory to a wide range of other work that adopts a normative approach to decision making (e.g., Anderson, 1990; Bogacz et al., 2006; Geisler, 2003; Niv et al., 2006). A particularly strong resonance is with work taking a normative perspective on action understanding (Bekkering et al., 2000; Csibra & Gyorgy, 2007; Gergely & Csibra, 2003), some of which has also taken an explicitly probabilistic approach (Baker, Saxe, & Tenenbaum, in press; Verma & Rao, 2006a).

Perhaps ironically, the work we have presented aligns less well with most existing work that explicitly addresses the role of “goals” in behavior. A great deal of work on goals focuses on the question of what specific classes of outcome humans are motivated to bring about (e.g., Deci & Ryan, 2000), a good example being the large literature on achievement motivation (Brunstein & Heckhausen, 2008). This issue is obviously orthogonal to the ones our theory addresses. However, one point that we have made that might inform studies of motivational content, is that behavior can be driven toward a specific outcome state through two fundamentally different mechanisms. In one, the ‘goal’ state operates as an explicit constraint on planning (as in the state query in our models). In the other, the goal state is not represented any differently from other states; what makes it, effectively, a goal state, is its association with a high reward value. Our simulations of the reward query show how overt behavior can tend consistently toward particular outcomes, without those outcomes ever being represented explicitly as goals. Interestingly, this aspect of the account may allow it to shed light on data suggesting that ‘goals’ can be implicit or unconscious (Bargh, Gollwitzer, Lee-Chai, Barndollar, & Trotschel, 2001).

Although the theory we have presented encompasses situations where goals are explicitly cued and adopted as hard constraints on subsequent decision making, it also asserts that, more often, goal selection is an integral part of goal-directed decision making, and furthermore that goal selection is accomplished, to some extent, in tandem with the processes that identify means to the end. This account thus differs, at least in its scope, from stage-like accounts which posit discrete goal selection and goal pursuit phases of planning (Lewin, Dembo, Festinger, & Sears, 1944; Oettingen & Gollwitzer, 2004).

The fact that our theory does not treat goals as necessarily fixed, pre-specified constraints on decision making sharply distinguishes it from many cognitive theories of planning and problem solving (Newell & Simon, 1971; Novick & Bassok, 2005; Unterrainer & Owen, 2006), as does our emphasis on the role of reward representations. Another important difference involves our focus on situations in which the decision maker has a good model of the causal and reward structure of the behavioral domain. Research into human problem solving has tended to focus on situations where the decision maker lacks an adequate and established model of the problem domain, or lacks the capacity to search such a model in sufficient depth. While this approach has led to important discoveries concerning bounds on rationality in decision making, and concerning heuristic decision-making procedures, it has tended to neglect what we consider to be the more common situation in everyday life, where adequate (if approximate) models are available to the decision maker, and where the search problem is tractable. From the standpoint of problem-solving research, such decision-making situations appear almost trivial; perhaps this is why the computational principles underlying them remain insufficiently studied.
Of course, ultimately what is needed is an account of goal-directed decision making that covers both ‘unproblematic’ decision making situations of the kind we have addressed here, and ‘problem’ situations involving inadequate models or insufficient representational or processing capacity. We consider below some ways in which the present theory might be developed, in order to better cover that continuum.

Many recent models of decision making have focused not on reward-based decision making, but on the formation of perceptual judgments (Bogacz et al., 2006; Gold & Shadlen, 2001, 2007). One might expect such models to invoke quite different principles from the ones involved in our framework. It is therefore rather surprising and interesting to note the existence of an underlying link. As mentioned earlier, most current theories of perceptual decision making are organized around sequential sampling, in which judgments evolve based on a repeated evaluation of perceptual evidence. The updating procedure shown in Equation 5 is mathematically equivalent to sequential sampling; the evolution of policies depicted in our simulation results is directly analogous to the evolving judgments involved in models of perceptual decision (in particular, models that portray interim judgments as based on the explicit computation of posterior probabilities; Beck & Pouget, 2007). This shared structure may prove useful in the effort to translate established models of perceptual decision making into the domain of reward-based choice (Rustichini, 2008). For example, it offers an explicit account for why reaction times in incentive choice are larger when options have similar utilities (see Rustichini, Dickhaut, Ghirardato, Smith, & Pardo, 2005), precisely analogous to prevailing explanations for the scaling of reaction times in perceptual discrimination tasks.

Relation to Neuroscience

Previous Computational Accounts.

As discussed in the Introduction, a number of neurobiologically inspired models have been advanced to address how goal-directed decision making — in most cases, path planning in spatial navigation — might be implemented in neural hardware (Hasselmo, 2005; Koene & Hasselmo, 2005; Muller et al., 1996; Reid & Staddon, 1998; Schmajuk & Thieme, 1992; Voicu & Schmajuk, 2002). The framework we have put forth shares a definite family resemblance with many of these models. As noted earlier, most neural models of planning posit a network of units representing environmental states or places, interconnected to form a ‘cognitive map’ of the environment. The propagation of activation within these networks clearly parallels the propagation of probability information through our model. A relative strength of our model is that it offers an explicit formal characterization of the computations involved in goal-directed decision making, establishing a link between these computations and inference-based computations in other information-processing domains, as well as to normative accounts of sequential decision making. By implementing goal-directed decision making in probabilistic terms, our models also naturally extend to settings involving uncertain outcomes and multiple sources of reward or cost, settings not generally dealt with in existing neural network models. Of course, given the relatively abstract computational approach our framework takes, it has less to say about the specific neuronal operations underlying goal-directed decision making. Nevertheless, we suspect it will be possible to translate between the formal account we have proposed and the more implementationally focused neural network models that have come before. One inviting approach to bridging between the two would be to leverage recent models implementing Bayesian inference procedures in neural terms (Beck & Pouget, 2007; Deneve, 2008; Litvak & Ullman, 2009; Ma et al., 2006; Pouget et al., 2003; Rao, 2006).
The present framework also shares close ties with a model proposed by Daw, Niv and Dayan (2005), which addressed the arbitration between goal-directed and habit systems in action selection. Unlike most neurobiologically oriented work in this area, the Daw, Niv and Dayan model framed goal-directed decision making in explicit formal terms, starting from a model-based reinforcement learning formulation. This permitted a rigorous, normative treatment of the decision process, including such aspects as decision uncertainty (a key element of their arbitration account). The present work is pitched at a similar level of abstraction, and also shares with the Daw model a focus on Bayesian inference. The framework we have presented makes more specific commitments concerning the nature of goal-directed decision making, placing probabilistic inference in a role different from the one it assumes in the Daw, Niv and Dayan (2005) model. On the other hand, the present work, unlike that of Daw, Niv and Dayan, does not yet address the role of learning or the interface between goal-directed and habitual action. As discussed below, these are two important areas for future development.

A notable aspect of the Daw, Niv and Dayan (2005) model is that it involves proposals concerning the neuroanatomic site of specific computational operations. The locus of goal-decision making, in that model, is prefrontal cortex. Other neural models of goal-directed decision making have ascribed the relevant operations to specific neuroanatomic sites or networks, including prefrontal cortex (Hasselmo, 2005), hippocampus (Muller et al., 1996), and basal ganglia (Girard et al., 2005). In the same spirit, our model is built around a specific set of hypotheses concerning functional localization. In contrast to some previous models, we have taken a ‘systems’ approach, portraying goal-directed decision as the result of a collaboration among several distinct neural systems, a policy system centering on DLPFC, a utility system centering on OFC and BLA, a state projection system centering on dorsal striatum and inferior frontal and medial temporal cortex, and an action system centering on motor cortices. Our theory views these systems as fitting together into a generative model, which relates behavioral policies and environmental conditions to actions, outcomes, and rewards. Goal-directed behavior is then viewed as involving several ways of querying this model and, in the most interesting case, inverting it through Bayes’ law.

We view it as important that this approach aligns with recent theories in visual neuroscience, in which the visual cortices are characterized as embodying a generative model for retinal inputs, and recognition is accomplished by inverting this model (Friston, 2005; Kersten, Mamassian, & Yuille, 2004; Lee & Mumford, 2003; Rao & Ballard, 1999). The relationship between this work and the ideas we have presented indicates the generality of the generative model approach, and the potential for integration that comes from approaching goal-directed decision making from this perspective. Interestingly, in very recent work, generative approaches have begun to appear in work on motor cortices (Kilner, Friston, & Frith, 2007; Wolpert, Doya, & Kawato, 2003). Going forward, it will be interesting to consider the relationship between this emerging research and the ideas we have advanced here.

Of course, beyond our overarching approach, our model attributes specific functions to the individual neural systems involved in goal-directed decision making. Having specified a set of computations these systems might collectively perform, we now revisit the evidence concerning their individual contributions to goal-directed decision making, concentrating on what we have termed the policy, reward, and state projection systems.

The Policy System and DLPFC.

Our model ascribes to the DLPFC the role of representing policies, mappings from states to actions. As noted earlier, this characterization fits neatly with existing accounts concerning the
role of DLPFC in cognitive control. The role of the DLPFC in our theory can, in fact, be seen as transposing into probabilistic terms the guided activation theory of Miller and Cohen (2001), or the information theoretic account of DLPFC function offered by Koechlin (2007). One way in which our theory goes beyond such accounts is that it provides an account of how policy representations are selected within DLPFC. There is surprisingly little previous work addressing this critical issue. One of the few exceptions is the model proposed by O'Reilly and Frank (2006), in which policy representations are ‘gated’ into prefrontal working memory circuits, through a mechanism shaped through dopamine-based reinforcement learning (see also Botvinick, Niv, & Barto, in press). In reinforcement-learning terminology, O'Reilly and Frank (2006) posit a ‘model-free’ mechanism for policy selection in DLPFC. This contrasts with the account we've proposed, in which prefrontal policy representations are selected based on a model supporting predictions concerning their consequences (see also Reynolds & Mozer, 2009). Of course, there may be no need to choose exclusively between model-based and model-free accounts, since as reviewed earlier, current evidence suggests that both mechanisms play a role in decision making.

Our view of DLPFC as carrying policy representations relates closely to work suggesting that the DLPFC codes for task sets or “rules” (Bunge et al., 2005; Wallis et al., 2001; White & Wise, 1999), to the extent that the latter can be viewed as mappings from environmental states to actions. This being said, there are also aspects of the task-set and rule accounts that go significantly beyond what is involved in our model, at least in its present stage of development. Perhaps most important is the issue of abstraction. Unlike the componential policy representations in our model, the rules or task sets addressed in studies of DLPFC tend to abstract over a set of stimulus-response associations. For example, a task representation for the Stroop task would code simultaneously for the mapping from red stimuli to the verbal response “red,” blue stimuli to the response “blue,” and so forth (Rougier, Noell, Braver, Cohen, & O'Reilly, 2005). This kind of “policy abstraction” (Botvinick et al., in press) suggests a hierarchical scheme for representing policies, and indeed recent research on DLPFC has given rise to the idea that this area may represent action policies at multiple hierarchical levels (Badre, 2008). In a later section, we discuss prospects for engaging the issue of abstraction and hierarchy in the present model.

As detailed earlier, one motivation for treating DLPFC as a hub for goal-directed decision making comes from rodent research, which has demonstrated an important role for the prelimbic cortex, an area that has been proposed to represent an analogue of the primate DLPFC. Unfortunately, it is not easy to evaluate the extent to which prelimbic cortical function fits with the policy-representation role it would be ascribed in our model. In fact, there is at least one challenging empirical finding. Ostlund and Balleine (2005) noted that devaluation sensitivity was disrupted by lesioning the prelimbic cortex prior to but not after instrumental learning. This seems to run counter to the notion that the prelimbic cortex is critical to goal-directed decision making, as it would be in our model. While we acknowledge the challenge implicit in this finding, we find it interesting to note that the Ostlund and Balleine experiments focused on choices between simple, non-sequential instrumental responses. Given that only one environmental state is relevant in such decision situations (i.e., the current environmental state), and this state is known at the time of the decision, the problem of policy selection reduces to that of action selection. Thus, in our model, goal-directed decision making under these circumstances could, in principle, be conducted without the policy variable, relying only on the action variable to distinguish among alternative line of behavior. Of course, even if this observation explains the persistence of devaluation sensitivity following prelimbic lesions, it does nothing to explain the disruptive effect of pre-training lesions. To engage such findings, our model will need to be
further developed to address learning. We consider the prospects for such an extension in a later section.

Given the role our model accords to the dorsal prefrontal cortex, the model gives rise to specific predictions about neural activity in this region. First, within the model, there is a tight coupling between policy representations in dorsal prefrontal cortex and representations in the state projection system. Thus, for example, Johnson and Redish (2007) have observed transient activation of hippocampal place cells during intervals of “deliberation,” as rodents pause at intersections in a maze. If, as Johnson and Redish suggest, these activations represent state projection, then according to our model place field activation should be accompanied by parallel activation in the prelimbic cortex of neurons coding for the plans of action necessary for bringing the animal to the relevant places.

Our model also makes predictions about the time-course of neural activation in the dorsal prefrontal cortex during goal-directed decision making. The reversal studied in Experiment 1 provides one example; the model predicts that, in animals facing analogous choice conditions, neurons in the dorsal prefrontal cortex should show a temporal profile of activation paralleling that shown in Figure 6B. The model also predicts that activity relating to multiple steps of future behavior should be present in the dorsal prefrontal cortex. Evidence in favor of this idea is already provided by neurophysiological studies in monkeys (e.g., Mushiake et al., 2006). The model makes finer-grained predictions in this setting, however. Specifically, the model predicts that the specific pattern of incentives prevailing during decision-making will affect the order in which decisions are made. For illustration, consider the three plots in Figure 6. Note that in one, the decision concerning the first move (at S\textsubscript{1}) evolves more quickly than the decision concerning the second (at S\textsubscript{2}). In another plot, reflecting a different incentive regime, the decision concerning the second move (at S\textsubscript{3}) evolves earlier than the S\textsubscript{1} decision. And in the third plot, the decisions evolve more or less in parallel. The model predicts that it should be possible to observe similar ordering effects at the level of single-unit activation in dorsal prefrontal cortex, under appropriate task circumstances.

The Reward System, OFC and BLA.

As reviewed earlier, the involvement of OFC and BLA in representing incentive value is well established. The present theory adds to existing work a specific account of the functional relations between the reward system, based in OFC and BLA, and other neural systems, which give rise to goal-directed decision making. In particular, our theory stipulates that the reward system exerts its influence not through direct connections with structures representing plans or actions (pace Koechlin & Hyafil, 2007; Wallis, 2007), but instead through connections with a state projection system, which in turn connects to structures representing action. This idea may provide a context for interpreting the dense connections running from OFC and BLA to medial temporal and medial frontal cortex, regions that have been tentatively linked with state projection. It may also provide a context for interpreting the convergence of projections from dorsal and orbital frontal cortex on the caudate nucleus (Cavada et al., 2000), another region implicated in state projection.

Another novel implication of our theory is that OFC and/or BLA should represent the incentive value not only of current states, but also of states projected under current multi-step plans. Neurophysiologic studies have begun to look at representations of multi-step plans in dorsolateral PFC (Mushiake et al., 2006; Saito et al., 2005). To our knowledge, no work has yet been done examining activity within OFC and BLA in this setting.
Existing empirical data present two challenges for the theory we have presented. First, recent experiments by Ostlund and Balleine (2007) led them to conclude that the OFC in rodents may be involved primarily in Pavlovian learning, being less involved in instrumental learning. If true, this would obviously conflict with the role our model accords to OFC. The proposal from Ostlund and Balleine (2007) contrasts with the conclusions drawn from other work on OFC, in particular work in primates, so implications for our model (and beyond) are not yet clear. A more obvious challenge for our model comes from its treatment of OFC and BLA as one simple functional unit. Although there is evidence that these two structures collaborate closely in the representation of incentive value (Baxter et al., 2000; Schoenbaum et al., 2003), there are also important functional differences between OFC and BLA (Arana et al., 2003; Corbit & Balleine, 2005) which our model does not yet address.

The State Projection System

Within our model, the state projection system is responsible for representing anticipated future states and for encoding the transition function, relating actions to outcomes. As reviewed earlier, recent research points toward medial temporal and inferior medial cortices, as well as dorsal striatum, as plausible substrates for this function. The framework we have presented offers an account of how these structures might collaborate with frontal regions responsible for representing policies and actions, and with OFC and BLA, to support goal-directed decision making. One prediction of this account is that activity in regions performing state projection should decline following damage to components of the reward system (e.g., OFC), action system (e.g., premotor cortex), or policy system (e.g., DLPFC). A more specific prediction derives from our Experiment 2, which provided behavioral evidence for optimistic state projections. Presumably, parallel effects should be observed at the neural level in brain structures coding for predicted future states.

Relation to Machine Learning Accounts

As acknowledged in the Introduction, there is a rich literature on goal-directed decision making (though under other names) in operations research, control theory and machine learning. These disciplines have generated a wide range of algorithms for maximizing expected reward or utility in multi-step problems. To some extent, however, such algorithms share an underlying resemblance to one canonical algorithm referred to as policy iteration. Policy iteration centers on computing a value function, which relates each environmental state to an estimate of the cumulative future reward that may be expected going forward from that step. After computing this value function, policy iteration updates the policy to favor actions leading to high-value states (see Puterman, 2005; Sutton & Barto, 1998).

On the surface, the computational approach adopted in the present paper appears rather different from policy iteration. In particular, the basic elements of the approach — policy, action, state and utility variables — do not include a direct analogue of state values. However, beneath the surface, the relationship to policy iteration is close. Indeed, Toussaint and Storkey (2006) analyzed an inference-based algorithm closely related (though not identical) to the one we have leveraged in the present work, and found it to be mathematically equivalent to policy iteration; state values were ‘hidden’ in the intermediate results involved in computing conditional probabilities.

It is important to note that Toussaint and Storkey (2006) go on to show that this equivalence holds only under certain restrictions, and when these are removed inference-based decision making may have computational advantages over traditional policy iteration (for example,
domains with partially observable or factored state). However, our intention here is not to argue for the superiority, in purely computational terms, of probabilistic inference as an approach to goal-directed decision making. Nor, in this sense, do we wish to highlight differences between the algorithm applied in the present work and other inference-based procedures that have been tried elsewhere (although meaningful differences do exist; see Botvinick & An, 2009). For psychology and neuroscience, we see the interest of the inference-based approach as deriving instead from two points emphasized in the Introduction: (1) Accounting for goal-directed decision making in these terms reveals the possibility that it may represent one domain-specific instance of a domain-general set of computational principles, which apply equally to perception, memory, motor control and elsewhere; (2) In the case of goal-directed decision making, as elsewhere, probabilistic inference provides a language equally relevant to characterizing cognitive and neural processes, allowing a transparent mapping between these two.

Having taken this rather ecumenical stance, it is nonetheless worth noting that different algorithms for goal-directed behavior do make different predictions. And, in fact, all do not cope equally well with the results of the two experiments we have reported. Experiment 1 is inconsistent with one simple but efficient, and widely applied, algorithm for sequential decision making, referred to as backward induction (see Puterman, 2005). Backward induction begins with a decision about what policy to follow in the last step of a multi-step task. This decision then provides the context for deciding what policy to adopt on the preceding step, and this process continues until the first step is reached. Contrary to backward induction, the results of Experiment 1 clearly show that, in human decision making, work can begin at early steps before the policy at later steps is finalized. The results of Experiment 2 are still more constraining, since they contradict not only backward induction but also policy iteration. Indeed, we are aware of no algorithmic approach other than the inference-based one we have focused on that would explain the kind of optimistic state prediction observed in Experiment 2.

**Directions for Further Development**

Obviously, the framework we have put forth provides only a rough sketch of the processes underlying goal-directed decision making. In laying out its details, we have acknowledged some unaddressed issues and potential challenges. In the present section, we confront some of these more directly, and point out some additional areas for further development.

On important limitation of our model, at least in the instantiations described in the present paper, is that it treats the transition and reward functions as given. In contrast, of course, human and animal decision makers must learn much of the transition and reward functions from experience: What will happen if I press this button on my new cell phone? How much do I like (or dislike) escargot? A full account of goal-directed decision making must address the learning processes that encode the answers to such questions, and the interaction of those processes with decision-making procedures. Fortunately, the theory we have presented is quite amenable to such an extension. Formal methods for learning in graphical models are well-developed (Jordan, 1998), and analogies have already been made between the relevant algorithms and learning processes in humans (Chater, Tenenbaum, & Yuille, 2006; Gopnik & Schulz, 2007). Previous work in machine learning also provides starting points for integrating such learning into inference-based models of goal-directed decision making (e.g. Attias, 2003).

One interesting issue that arises when learning and action selection are combined is that action choices can affect what is learned. The learner can thus engage in ‘active learning,’ in which actions are taken to maximize information gain (Castro et al., 2008; Kruschke, 2008; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Another setting where action can be motivated by
the ‘value of information’ is partial observability, where the state of the environment is not entirely available to immediate perception (Behrens, Woolrich, Walton, & Rushworth, 2007; Howard, 1966). The models we presented assumed full state observability. However, Toussaint (2006) and Furmston and Barber (2009) have described how similar principles can be applied to partially observable problems. Evaluating the fit between the resulting account and human behavior in analogous task contexts presents an interesting challenge.

Another, more immediate challenge for the theory we have presented is to take account of the rather strict capacity limitations attaching to human information processing. Such capacity limitations are glaringly evident in goal-directed decision making, as amply demonstrated by decades of research on human planning and problem solving. A central take-home message from such research is that human planners are incapable of reasoning precisely about complex problems, due largely to limitations on working memory capacity, and thus resort to a number of simpler problem-solving heuristics (Newell & Simon, 1971; Novick & Bassok, 2005; Unterrainer & Owen, 2006). How are the relevant limitations to be understood within the framework we have proposed? One lead is presented by recent work in other behavioral domains, suggesting that human cognition does not leverage probability distributions in their entirety, but rather only samples from such distributions. Under this approach, information processing capacity limitations are understood as limitations on the number of samples that can be made during a single decision-making event. This general idea, which leverages machine learning algorithms for approximate inference, has been applied to magnitude estimation (Vul & Pashler, 2008) and sentence processing (Levy, Reali, & Griffiths, 2009). Investigating its applicability to goal-directed decision making is a challenge we are currently pursuing.

One way of mitigating the impact of limited capacity on goal-directed decision making, both in machine learning and in human cognition, is through hierarchical representation. Hierarchical action representations simplify the planning problem, allowing plans to reach deeper into the future through efficient coding of action sub-sequences (see Botvinick et al., in press). As discussed in Simulation 2B, Ostlund and colleagues (2009) reported devaluation behavior which they interpreted as direct evidence for “chunked” action representations in goal-directed behavior. Although, in our earlier discussion, we suggested the relevant data might be explained without chunking, it seems certain that, in the general case, hierarchical action representations do play a role in goal directed decision making. In recent work, Toussaint and colleagues (Toussaint, Charlin, & Poupart, 2008) have provided an initial demonstration of how hierarchical representation can be integrated with inference-based planning. It would be interesting to consider how the relevant computational issues relate to recent findings suggesting that prefrontal cortex houses a topographically organized hierarchy of action representations (Badre, 2008).

In the present work, we have modeled goal-directed decision making in isolation, but as recent work has emphasized, human and animal behavior also rests upon habitual action selection, supported by different computational and neural mechanisms. A final important area for further development of the current account involves the question of how goal-directed decision making mechanisms interface with the habit system (Botvinick & Plaut, 2006; Cooper & Shallice, 2006; Coutureau & Killcross, 2003; Daw et al., 2005; Killcross & Coutureau, 2003). One way of modeling the role of habits in the present framework might be as additional inputs to policy variables, biasing policy selection toward habitual configurations. How well this approach aligns with available behavioral and neural evidence concerning the goal/habit interface will be an interesting question to pursue.
Conclusion

Our title promised an “integrative” account of goal-directed decision making. The theory we have presented fits that description in three ways. First, it embraces several qualitatively different forms of goal-directed decision making, fitting these into a unifying framework. Second, by describing goal-directed decision making in terms of probabilistic inference, our account aligns it with information processing in numerous other domains (e.g., memory, perception and motor control), which have also recently come to be understood in probabilistic terms. Goal-directed decision making is thus portrayed as simply one particular instantiation of a more general set of computational principles, rather than a domain involving special-purpose mechanisms. Finally, our account attempts to span between cognitive and neural levels of description, relying on computation as a Rosetta stone for translating between the two.

Of course, the fact that a theory is integrative provides no guarantee that it is correct. At the very least, therefore, we hope the present work will provide a useful survey marker, as further work is conducted toward charting out this important domain.
References


Botvinick and colleagues

Goal-Directed Decision Making


Appendix

The main text introduced a recursive procedure for solving Markov decision problems within graphical models of the kind displayed in Figures 2 and 4. Here we provide formal proofs of monotonicity and convergence (based on Botvinick & An, 2009), which guarantee that the algorithm will converge to an optimal policy. To recap, the procedure is as follows: (1) Initialize the policy nodes with any set of non-deterministic priors. (2) Treating the initial state and \( u_G \) as observed variables (\( u_G = 1 \)), use standard belief propagation or a comparable algorithm to infer the posterior distributions over all policy nodes. (3) Set the prior distributions over the policy nodes to the values (posteriors) obtained in step 2. (4) Go to step 2. The proofs follow:

**Monotonicity**

We show first that, at each policy node, the probability associated with the optimal policy will rise on every iteration. Define \( \pi^* \) as follows:

\[
p(u_G | \pi^*, \pi') > p(u_G | \pi', \pi^*), \forall \pi' \neq \pi^*
\]

where \( \pi^* \) is the current set of probability distributions at all policy nodes on subsequent time-steps. (Note that we assume here, for simplicity, that there is a unique optimal policy.) The objective is to establish that:

\[
p(\pi^*_n) > p(\pi^*_{n-1})
\]

where \( n \) indexes processing iterations. The dynamics of the network entail that

\[
p(\pi_n) = p(\pi_{n-1} | u_G)
\]

where \( \pi \) represents any value (i.e., policy) of the decision node being considered. Substituting this into A2 gives

\[
p(\pi^*_{n-1} | u_G) > p(\pi^*_{n-1})
\]

From this point on the focus is on a single iteration, which permits us to omit the relevant subscripts. Applying Bayes’ law to A4 yields

\[
\frac{p(u_G | \pi^*) p(\pi^*)}{\sum_{\pi} p(u_G | \pi) p(\pi)} > p(\pi^*)
\]

Canceling, and bringing the denominator up, this becomes

\[
p(u_G | \pi^*) > \sum_{\pi} p(u_G | \pi) p(\pi)
\]
Rewriting the left hand side, we obtain

\[ \sum_{\pi} p\left(u_G | \pi^*\right) p(\pi) > \sum_{\pi} p\left(u_G | \pi\right) p(\pi) \quad (A7) \]

Subtracting and further rearranging:

\[ \sum_{\pi} \left[ p\left(u_G | \pi^*\right) - p\left(u_G | \pi\right) \right] p(\pi) > 0 \quad (A8) \]

\[ \left[ p\left(u_G | \pi^*\right) - p\left(u_G | \pi^*\right) \right] p(\pi^*) + \sum_{\pi \neq \pi^*} \left[ p\left(u_G | \pi^*\right) - p\left(u_G | \pi^*\right) \right] p(\pi^*) > 0 \quad (A9) \]

\[ \sum_{\pi \neq \pi^*} \left[ p\left(u_G | \pi^*\right) - p\left(u_G | \pi^*\right) \right] p(\pi^*) > 0 \quad (A10) \]

Note that this last inequality (A10) follows from the definition of \( \pi^* \).

**Remark:** Of course, the identity of \( \pi^* \) depends on \( \pi^* \). In particular, the policy \( \pi^* \) will only be part of a globally optimal plan if the set of choices \( \pi^* \) is optimal. Fortunately, this requirement is guaranteed to be satisfied, as long as no upper bound is placed on the number of processing cycles. Recalling that we are considering only finite-horizon problems, note that for policies leading to states with no successors, \( \pi^* \) is empty. Thus \( \pi^* \) at the relevant policy nodes is fixed, and is guaranteed to be part of the optimal policy. The proof above shows that \( \pi^* \) will continuously rise. Once it reaches a maximum, \( \pi^* \) at immediately preceding decisions will perforce fit with the globally optimal policy. The process works backward, in the fashion of backward induction.

**Convergence**

Continuing with the same notation, we show now that

\[ \lim_{n \to \infty} p_n\left( \pi^* | u_G \right) = 1 \quad (A11) \]

Note that, if we apply Bayes’ law recursively,

\[ p_n\left( \pi^* | u_G \right) = \frac{p\left(u_G | \pi^*\right) p_n\left( \pi^* \right)}{p_n\left(u_G \right)} = \frac{p\left(u_G | \pi^*\right)^2 p_{n-1}\left( \pi^* \right)}{p_n\left(u_G \right) p_{n-1}\left(u_G \right)} = \frac{p\left(u_G | \pi^*\right)^3 p_{n-2}\left( \pi^* \right)}{p_n\left(u_G \right) p_{n-1}\left(u_G \right) p_{n-2}\left(u_G \right)} \ldots \quad (A12) \]

Thus,

\[ p_n\left( \pi^* | u_G \right) = \frac{p\left(u_G | \pi^*\right)^n p_1 \left( \pi^* \right)}{\prod_{m=1}^{n} p_m\left(u_G \right)} \quad (A13) \]
Therefore, what we wish to prove is

\[
\frac{p(u_G | \pi^*)}{\prod_{n=1}^{\infty} p_i(u_G)} p_1(\pi^*) = 1
\]  

(A14)

or, rearranging,

\[
\prod_{n=1}^{\infty} \frac{p_n(u_G)}{p(u_G | \pi^*)} = p_1(\pi^*).
\]  

(A15)

Note that, given the stipulated relationship between \( p(\pi) \) on each processing iteration and \( p(\pi | u_G) \) on the previous iteration,

\[
p_n(u_G) = \sum_\pi p(u_G | \pi)p_n(\pi) = \sum_\pi p(u_G | \pi)p_{n-1}(\pi | u_G) = \sum_\pi \frac{p(u_G | \pi)^2 p_{n-1}(\pi)}{p_{n-1}(u_G)}
\]

\[
= \frac{\sum_\pi p(u_G | \pi)^3 p_{n-1}(\pi)}{p_{n-1}(u_G) p_{n-2}(u_G)} = \frac{\sum_\pi p(u_G | \pi)^{n+1} p_{n-1}(\pi)}{p_{n-1}(u_G) p_{n-2}(u_G) p_{n-3}(u_G) \ldots}
\]  

(A16)

With this in mind, we can rewrite the left hand side product in A15 as follows:

\[
\frac{p_i(u_G)}{p(u_G | \pi^*) p_1(u_G)} \cdot \frac{\sum_\pi p(u_G | \pi)^2 p_1(\pi)}{p(u_G | \pi^*) p_1(u_G)} \cdot \frac{\sum_\pi p(u_G | \pi)^3 p_1(\pi)}{p(u_G | \pi^*) p_1(u_G) p_2(u_G)} \cdot \frac{\sum_\pi p(u_G | \pi)^4 p_1(\pi)}{p(u_G | \pi^*) p_1(u_G) p_2(u_G) p_3(u_G) \ldots}
\]  

(A17)

Note that, given A16, the numerator in each factor of A17 cancels with the denominator in the subsequent factor, leaving only \( p(u_G | \pi^*) \) in that denominator. The expression can thus be rewritten as

\[
\frac{1}{p(u_G | \pi^*)} \cdot \frac{1}{p(u_G | \pi^*)} \cdot \frac{1}{p(u_G | \pi^*)} \cdot \frac{1}{p(u_G | \pi^*)} \ldots \frac{\sum_\pi p(u_G | \pi)^\infty p_1(\pi)}{p(u_G | \pi^*)} = \sum_\pi \frac{p(u_G | \pi)^\infty p_1(\pi)}{p(u_G | \pi^*)}.
\]  

(A18)

The objective is then to show that the above equals \( p(\pi^*) \). It proceeds directly from the definition of \( \pi^* \) that, for all \( \pi \) other than \( \pi^* \),

\[
\frac{p(u_G | \pi)}{p(u_G | \pi^*)} < 1
\]  

(A19)
Thus, all but one of the terms in the sum above approach zero, and the remaining term equals \( p(\pi^*) \). Thus,

\[
\sum_{\pi} \frac{p(u_c | \pi)}{p(\pi)} p_1(\pi) = p_1(\pi^*) \quad \therefore
given (A20)
Author Note

Thanks to James An for contributions to the simulation work, and to David Blei, Peter Dayan, Nathaniel Daw, Andrew Ledvina, Yael Niv, Francisco Pereira, Mike Todd, and participants in the Neural Basis of Goal Directed Behavior seminar at Princeton for useful discussion. The present research was supported by a grant from the McDonnell Foundation.
Notes

1. Technically speaking, there is a distinction between Bayesian networks and causal networks, which plays a significant role in the work of Gopnik and colleagues (2004). Though this distinction is critical in some contexts, it does not interact strongly with the ideas we present here, and in view of this we shall speak equivalently of Bayesian and causal networks, and of conditional probabilities and causal dependencies.

2. Representing this as a multinomial variable is obviously a massive simplification. Graphical model formalism can accommodate state represented in a factored or distributed way, or even continuous. The same applies to the action representations discussed below.

3. We shall use the term *query* to refer to this entire operation, and sometimes to the variable assumed to be known. This is a little different from the conventional usage, where the variable one is asking about is referred to as the 'query variable.'

4. Here and throughout the paper, we simplify notation by simply writing a variable name to indicate a specific assignment of the variable.

5. Adding the observation ‘not punished’ to the state query might help. However, for this to be a satisfying explanation for the likely behavior, one would require an account of how the child came to adopt this as part of the goal state.

6. The approach is also closely related to the Expectation-Maximization or EM algorithm, which has also been used to allow solution of Markov decision problems using probabilistic inference (Dayan & Hinton, 1997; Hoffman et al., 2009; Toussaint & Storkey, 2006).

7. Although we have focused on OFC and BLA as substrates for the representation of utility, it should be noted that there is evidence that the costs of effort may be represented in different structures, in particular the dorsal anterior cingulate cortex (Botvinick, Huffstetler, & McGuire, 2009; Rudebeck, Walton, Smyth, Bannerman, & Rushworth, 2006).

8. Note that the rat’s ‘state’ at $S_2$ and $S_3$ might thus be understood as factoring in an internal representation of past actions. However, as Ostlund and colleagues (2009) note, this is not strictly necessary, since visual, tactile, and proprioceptive information might suffice to discriminate among the relevant situations.

9. Ostlund et al. (2009) base their interpretation not only on the performance of normal rats, but also on that of animals with lesions to the dorsomedial prefrontal cortex. They propose, in particular, that such lesions disrupt action ‘chunk’ representations. However, Ostlund and colleagues (2009) briefly acknowledge an alternative explanation for their data, which is that dorsomedial prefrontal lesions reduce the number of future steps animals are able to consider during planning, making them “relatively myopic in their evaluation of action–outcome relationships” (p. 8285).
Table 1. Stimulus selection in Experiment 1. Point values are indicated for characters in each box. Assignments of Box 1 and 2 to left and right positions varied randomly across trials.

<table>
<thead>
<tr>
<th>Box 1</th>
<th>Box 2</th>
<th>Trial Type</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 9 8 3</td>
<td>Match</td>
<td>0.03125</td>
<td></td>
</tr>
<tr>
<td>10 8 9 3</td>
<td>Match</td>
<td>0.03125</td>
<td></td>
</tr>
<tr>
<td>10 9 8 0</td>
<td>Match</td>
<td>0.03125</td>
<td></td>
</tr>
<tr>
<td>10 8 9 0</td>
<td>Match</td>
<td>0.03125</td>
<td></td>
</tr>
<tr>
<td>10 9 3 0</td>
<td>Match</td>
<td>0.03125</td>
<td></td>
</tr>
<tr>
<td>10 3 9 0</td>
<td>Match</td>
<td>0.03125</td>
<td></td>
</tr>
<tr>
<td>10 8 3 0</td>
<td>Match</td>
<td>0.03125</td>
<td></td>
</tr>
<tr>
<td>10 3 8 0</td>
<td>Match</td>
<td>0.03125</td>
<td></td>
</tr>
<tr>
<td>9 8 3 0</td>
<td>Match</td>
<td>0.125</td>
<td></td>
</tr>
<tr>
<td>9 3 8 0</td>
<td>Match</td>
<td>0.125</td>
<td></td>
</tr>
<tr>
<td>10 3 9 8</td>
<td>Mismatch</td>
<td>0.0625</td>
<td></td>
</tr>
<tr>
<td>10 0 9 8</td>
<td>Mismatch</td>
<td>0.0625</td>
<td></td>
</tr>
<tr>
<td>10 0 9 3</td>
<td>Mismatch</td>
<td>0.0625</td>
<td></td>
</tr>
<tr>
<td>10 0 8 3</td>
<td>Mismatch</td>
<td>0.0625</td>
<td></td>
</tr>
<tr>
<td>9 0 8 3</td>
<td>Mismatch</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>
Figure Captions

1. Left: Maze used to demonstrate latent learning, redrawn from Blodgett (1929, page 117). D = door. Right: Maze used to demonstrate detour behavior, redrawn from Tolman and Honzik (1930, page 223).

2. Elements of the computational account.

3 A graphical model of the earthquake story, adapted from Pearl (1988, p. 9).

4. Three forms of query, in the cases of single-step (left) and multi-step (right) decision.

5. Results of simulation studies 1A and 1B. Blue and green traces indicate the posterior probability of left and right actions at each processing iteration. Red traces indicate the probability of p(u = 1) given the mixture of policies at each iteration, proportional to the expected reward for that mixture. Dashed red lines indicate p(u = 1) for the optimal policy. A-C: Simulation 1A. A: Simple instrumental choice. B: Devaluation. C: Contingency degradation. D: Results of Simulation 1B.

6. A: T-maze scenario from Niv, Joel and Dayan (2006). B-D: Results of Simulation 2A, formatted as in Figure 5. B: Choice under the incentive values indicated in panel A. C: Motivational change. D: Effort-based decision.

7. Results of Simulation 2B. Green and blue curves indicate the probability of executing the two-step sequences left-right and right-left, from among all (four) possible two-step sequences. The two most central data series are offset for legibility; the values were in fact precisely equal across the two.

8. Results from Simulation 2C. A: Discretization of the maze from Tolman and Honzik (1930) into nine locations or states. B: Choice behavior with no obstruction (including expected values as in Figures 5 and 6). C: With obstruction at location a in panel A. D: With obstruction at location b.

9. Results from Simulation 2D, indicating the discretization of the maze from Blodgett (1929) into 13 states or locations. Heavy arrows indicate the path chosen following introduction of reward at S13.

10. Experiment 1. A: Example display. B: Data from Gratton et al. (1988, p. 335). C: Simulation of box selection in the experiment. Curves indicate the posterior probability of selecting the correct box over processing iterations, for match and mismatch trials, using the reward values shown in the insets. D: Experimental results. Bars indicate standard error of the mean.

11. Experiment 2. A: Example display. B: Proportion of trials involving choice of each character. Bars in this and subsequent panels indicate standard error. C: Mean visual analogue scale ratings. D: Mean ratings for the most valuable character, according to the value of the unchosen character.
Policy System
- Dorsolateral PFC
- Premotor, VLPFC, BA 10

Action System
- Premotor cortex, SMA

State Projection System
- Medial temporal cortex
- Medial frontal / parietal cortex
- Caudate nucleus

Reward System
- Orbitofrontal cortex
- Basolateral amygdala