Forum

Planning as inference

Matthew Botvinick[1] and Marc Toussaint[2]

[1] Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ, USA
[2] Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany

Recent developments in decision-making research are bringing the topic of planning back to center stage in cognitive science. This renewed interest reopens an old, but still unanswered question: how exactly does planning happen? What are the underlying information processing operations and how are they implemented in the brain? Although a range of interesting possibilities exists, recent work has introduced a potentially transformative new idea, according to which planning is accomplished through probabilistic inference.

Behavioral and neuroscientific data on reward-based decision making increasingly point to a fundamental distinction between habitual and goal-directed action selection. Habits, in this context, are actions arising from direct situation-response associations. Goal-directed action, in contrast, involves prospective planning: selection among actions based on a forecast of their potential outcomes [1,2].

Between these two forms of decision-making, much more is presently known concerning habit. Here, abundant evidence supports the relevance of temporal-difference procedures from reinforcement learning: dopaminergic inputs to the striatum appear to convey a reward-prediction error signal, which drives adaptive updates in striatal representations of state value and (habitual) action preference [1].

Regrettably, there is no corresponding story for goal-directed decision making. Recent lesion and neuroimaging work does provide important clues about localization, implicating segments of prefrontal cortex and dorsal striatum [1,2]. However, characterizing the actual information-processing operations that underlie goal-directed action selection remains an unresolved problem. A growing awareness of this challenge has been drawing the problem of planning back to center stage in cognitive science.

Traditional perspectives on planning

So how might the brain accomplish planning? In psychology, the classical approach to this question focuses on planning problems involving a specific *a priori* goal. Although the study of such tasks has yielded important insights, it stops short of the more general problem, which centers on the generic goal of reward maximization. The classical approach also concentrates on cases where action outcomes are perfectly predictable, something that is not characteristic of most real life settings. Ultimately, what we need is an account of how human decision-makers solve 'Markov decision problems': given a set of potential situations or states, a set of available actions, a set of probabilistic action-outcome relationships, and a set of preferences over outcomes, how is a plan of action cobbled together?

The field of operations research, which centers on this question, offers a number of candidate procedures. In dynamic programming, each state is associated with a value (a prediction of future reward), which is set through a repeated exchange of information between adjacent states. Actions are then selected by aiming for outcomes with high value. Model-based reinforcement learning algorithms, as applied in recent work on planning [3], do the same thing, but by chaining forward from the decision-maker's current state, effecting a kind of 'tree search' that focuses computational effort on reachable states.

Like classical cognitive models, dynamic programming and model-based reinforcement learning models offer a critical lever for uncovering the mechanisms that underlie human planning. However, we believe that additional, and possibly decisive, leverage may be offered by a third perspective on planning, which has only recently been crystallizing. This perspective reconceptualizes planning as a matter of probabilistic inference.

Planning as inference: the basic idea

Under the planning-as-inference (PAI) view, the decision-making agent makes use of an internal cognitive model, which represents the future as a joint probability distribution over actions, outcome states, and rewards (Figure 1a, b). This generative model allows the agent to attach a probability to any potential action-outcome-reward sequence.

To plan, the agent can use its internal model to sample potential action-outcome trajectories, essentially using it to perform tree search. However, because the model specifies a probability distribution, the agent also has another option: it can condition on reward, that is, the agent can start from the initial assumption that its actions will yield reward, and then use inverse inference to discover the actions that render this assumption most plausible. In slightly more technical terms, one can view the agent's strategy or action policy as a set of parameters specifying, for each state, a probability distribution over actions. Under PAI, the agent uses probabilistic inference to discover the maximum likelihood values for these parameters, conditional on future reward (Figure 1a).

PAI in machine learning and robotics

The seeds of PAI were planted in artificial intelligence and machine learning research as early as the 1980s (for this background, see [4,5]). Since then, evolving techniques for

Corresponding author: Botvinick, M. (matthewb@princeton.edu)

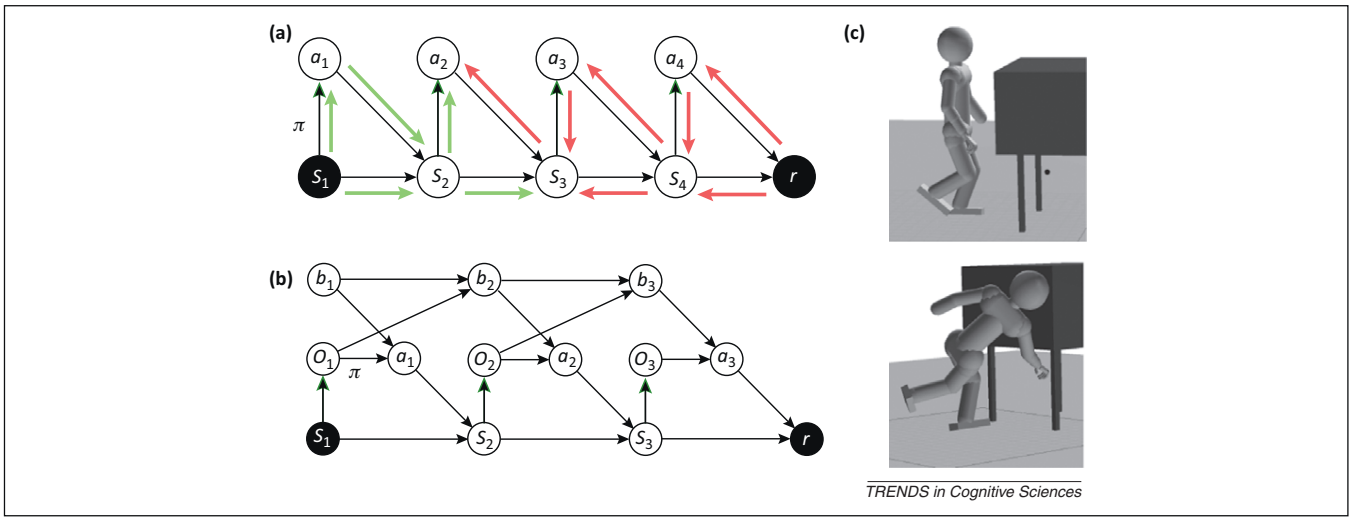

Figure 1. Planning as inference (PAI). **(a)** Schematic illustration of one implementation of PAI. The decision-maker's internal generative model is represented as a Bayesian network. Nodes correspond to random variables representing states (*s*) and actions (*a*) at successive time-steps (subscripts), as well as ultimate reward (*r*). Black arrows indicate conditional dependencies, with distributions over receiving variables depending on sending variables. Arrows connecting state to action variables carry the action policy (π). Shading of a node indicates that the value of the corresponding variable is known or stipulated. PAI assumes that reward is received and searches for the policy that is most probable under this assumption. This can be accomplished through procedures in which 'messages' are passed between variables. In the present implementation, messages propagating forward from the initial state (green) indicate the probability of encountering specific successor states. Messages propagating backward from reward (red) indicate the likelihood of having visited specific states, given the assumption of reward receipt. Forward and backward messages are integrated to yield probabilities of specific state-action pairs, and these form the basis for an updated policy. If this inference process is iteratively repeated, the model will converge to an optimal (reward-maximizing) policy. A straightforward elaboration of the model depicted here allows PAI to accommodate the fact that rewards can occur at any time and to encompass 'infinite horizon' scenarios where there is no fixed termination step. Adapted from [4]. **(b)** The graphical model representation can be extended to a variety of problem settings, allowing PAI to be flexibly applied. The model shown here captures the structure of a 'partially observable Markov decision problem', in which the agent receives (potentially ambiguous) observations (*o*) from the environment, rather than having direct access to its true underlying state (*s*). The agent maintains an internal memory or belief state (*b*) based on the history of its actions and observations, which in turn provides a basis for action selection. **(c)** In robotics, PAI has been applied to difficult motor control problems. In this scenario, approximate inference discovers a feasible posture for grasping an occluded target. Reproduced, with permission, from [7].

PAI have been increasingly applied to support planning in artificial agents and to solve stochastic optimal control problems in robotics. In both of these settings, recent implementations of PAI have yielded computational benefits over traditional techniques, discovering optimal solutions more quickly and, in some cases, tackling complex problems that otherwise appeared intractable (see, e.g., [6,7]; Figure 1).

Behind these successes lie two critical developments in PAI theory. First, increasing precision and generality has been gained in understanding the basic computational problem. In contemporary formulations, PAI is framed as involving a minimization of the (Kullback-Leibler) divergence between two probability distributions: the marginal distribution over states and actions under the agent's current policy, and the corresponding posterior distribution, under the assumption of future reward [6,8]. (This is complementary to previous formulations which introduce a mixture model with the time of reward as a random variable [4]). Second, on the algorithmic front, increasingly efficient methods have been developed for accomplishing the pivotal minimization, building on general procedures for parameter estimation (in particular, the expectation-maximization (EM) algorithm [4,6,8]).

On a coarse level, the resulting approach involves two phases. The first centers on computing or estimating the key probability distributions. In recent work, this has often been accomplished through 'message passing' within the underlying generative *model*. Here, marginal distributions for each variable are computed based on a local exchange of information among small subsets of variables within the larger model (Figure 1a). The second step involves updating the agent's policy, so as to bring its distribution of behaviors closer to the target distribution. Cycling between these two steps yields an iterative procedure, which gradually hones in on an optimal plan of action.

Implications for psychology and neuroscience

As cognitive and neuroscientific research has reengaged with the topic of planning, recent work has begun to explore the potential relevance of PAI [5,9]. Although still at its inception, such work already makes clear why PAI may hold special interest.

One appealing aspect of PAI is that it brings planning under the same umbrella as other forms of information processing. There has been a recent surge of interest in the idea that essentially all cognitive and neural computation can be understood in terms of probabilistic inference. Inference-based analyses have become central in areas ranging from perception and motor control to language processing and social cognition [10]. PAI offers a way of bringing planning within the same unifying view, characterizing this apparently special domain in domain-general terms.

On a more detailed level, the pivotal role that PAI accords to model inversion brings planning into close contact with Bayesian theories of visual perception, where inverse inference has been a key motif [10]. The analogy has been highlighted vividly in work by Friston and colleagues, where a common set of ideas, centering in part on model inversion, has been applied across a diverse set of domains including planning [9].

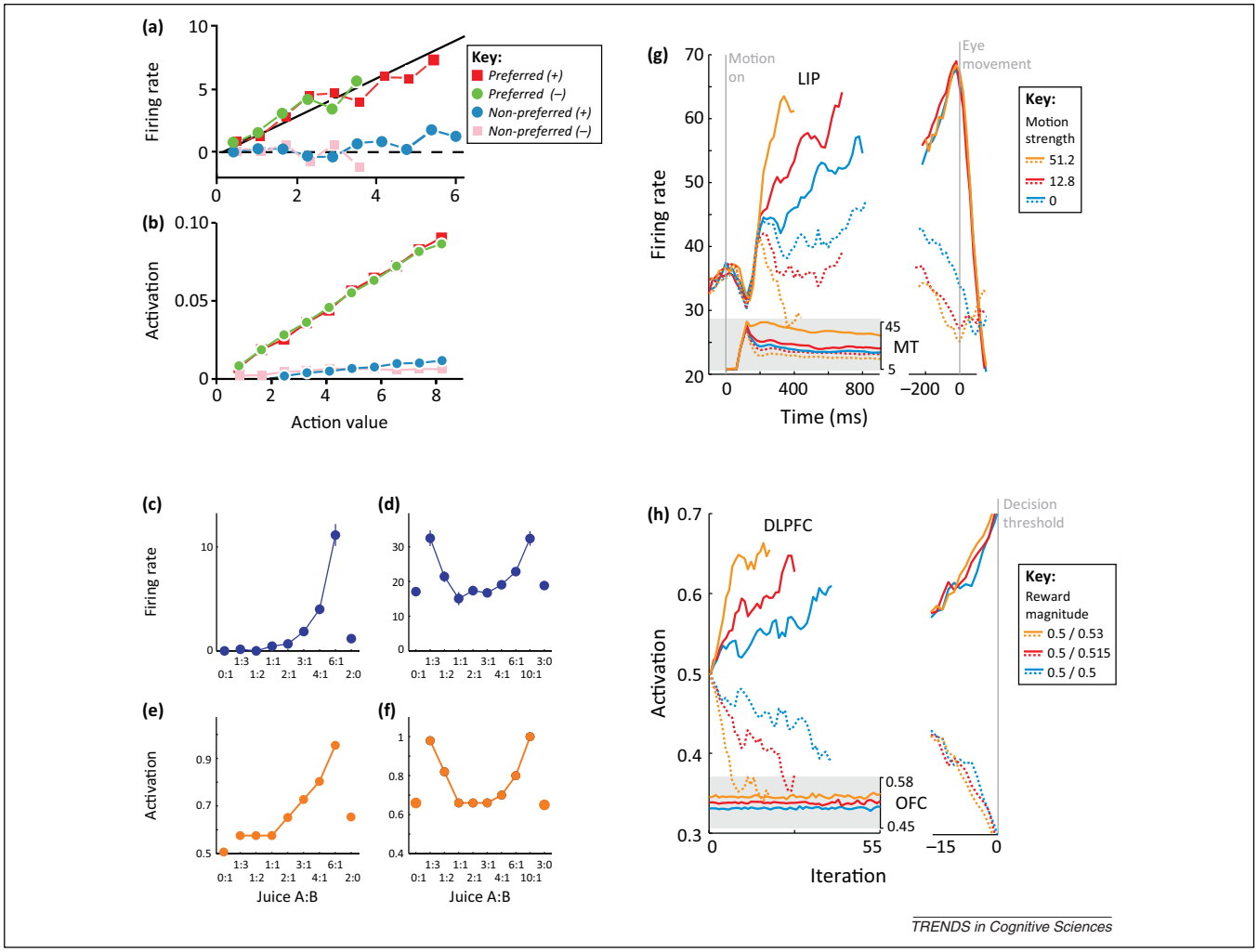

Figure 2. Neurophysiological findings paired with simulations from Solway and Botvinick [5]. **(a)** Single-unit recording data from [14]. In this study, monkeys chose between two visual targets yielding different quantities of juice. Neurons in the dorsal striatum coded for action *value:* individual neurons coded for specific (left vs right) eye movements, but with a firing rate that scaled with the reward to be expected for executing them. Firing rates were insensitive to the value of the opposing action. Preferred (non-preferred): portion of firing rate attributable to the value of a neuron's preferred (non-preferred) action. +/–: data from trials on which preferred (+) and non-preferred (–) action was executed. Adapted, with permission, from [14]. **(b)** Simulation results from [5]. When a neural network implementation of planning-as-inference (PAI) was applied to the task from [14], a subset of units tracked action value. Adapted, with permission, from [5]. **(c-d)** Neurophysiological data from [15]. Here, monkeys chose between different quantities and types of juice by making a saccade to one of two locations. Single-unit recordings in orbitofrontal cortex revealed neurons that coded for offer *value* (the utility of one of the juice offers, regardless of the competing option; panel c), and chosen value, the value of the option ultimately selected by the animal (panel d). Juice A:B: proffered units of each juice type (juice 'A' and juice 'B'). Adapted from [15]. **(e–f)** Simulation results from [5]. When a neural network implementation of PAI was applied the task from [15], a subset of units tracked offer value (panel e), while another subset tracked chosen value (panel f). Adapted from [5]. **(g)** Single-unit recording data from lateral intraparietal area (LIP) and middle temporal area (MT) during perceptual decision making (motion discrimination). MT carries a representation of momentary evidence concerning motion direction, whereas the ramping activity in LIP reflects information integration, akin to that involved in drift-diffusion decision models. A response is triggered when activity in LIP reaches a fixed threshold. Solid traces show data from trials where motion direction corresponded to the preferred direction of recorded neurons, dotted traces the anti-preferred direction. Adapted, with permission, from [11] © Annual Reviews Inc. **(h)** Results from a neural network implementation of PAI, showing how reward-based decision making might arise from processes analogous to those underlying perceptual decision making. Units proposed to capture a function of orbitofrontal cortex (OFC) code for momentary evidence concerning available rewards (as in panel g). Units representing action policies – modeling the role of the dorsolateral prefrontal cortex (DLPFC) – act as integrators, rising to a response threshold. Solid traces relate to units representing the chosen action policy, dotted traces to units representing the unchosen policy. Adapted from [5].

If PAI is relevant to the neural processes underlying goal-directed behavior, then one should expect to find that the brain implements a generative model linking plans to actions, actions to outcomes and outcomes to rewards (Figure 1). Pursuing this idea, Solway and Botvinick [5] pointed out an apparent correspondence between the components of the generative model in PAI and a specific set of neuroanatomic structures. To further explore this mapping, Solway and Botvinick [5] implemented a neural network model of the processes involved in PAI, leveraging recent work showing how neural computation may approximate message-passing procedures for probabilistic inference [10]. Individual units within the resulting neural network turned out to display response profiles closely matching those of individual prefrontal and striatal neurons recorded during reward-based decision making, as illustrated in Figure 2.

One surprising result in this work relates to the role of iteration. As noted earlier, PAI algorithms almost universally involve iterative processing, with repeated cycles of inference, each feeding into the next. Solway and Botvinick [5] pointed out the formal similarity between the role of iterative inference in PAI and its role in contemporary drift-diffusion models of perceptual decision-making [11],

identifying conditions under which the two are in fact equivalent. At a neural level, the iterative processes involved in perceptual decision making have been linked to specific substrates: research on perceptual choice shows parietal and frontal neurons acting as 'integrators', summing over sequential inputs from other areas that represent momentary evidence ([11]; Figure 2g). Simulations reported by Solway and Botvinick [5] indicate how computationally identical information-integration processes, playing out in prefrontal cortex, might support goal-directed decision making based on reward (Figure 2h).

Present opportunities and challenges

PAI appears to offer a promising new perspective on the time-honored problem of planning – one that reveals underlying commonalities with other cognitive functions and which may shed new light on relevant neural processes. Of course, a great deal of additional research will be needed if the apparent relevance of PAI to cognition and neural function is to be properly validated. To date, most work on PAI has been theoretical. A necessary next step will be to identify and test empirical predictions arising from PAI, predictions that differentiate the framework from other computational accounts. The recent surge in innovative experimental work on planning, both in animal learning and in behavioral and cognitive neuroscience, promises to provide a fertile context for this next stage of research.

A more specific challenge, which existing work on PAI has not fully engaged, derives from classical cognitive research on planning. A central take-home message from such research is that planning occurs under strict capacity limitations. Fully rational planning, based on an exhaustive, exact evaluation of all possible action-outcome trajectories, is infeasible. Instead, human planners display bounded rationality, applying heuristics and other simplifying strategies to obtain plans that 'satisfice'.

When difficult inference problems arise in machine learning, they can often be conquered through approximation techniques. In one major class of such techniques, precise representations of probability are substituted with estimates based on sampling. It has been proposed recently that such sampling-based approximations may be relevant to understanding human information processing under capacity limitations [12], as well as to stochastic operations within biological neural networks [13]. Importing the same idea into the domain of planning, by way of PAI, may offer a new way of understanding and modeling bounded rationality in planning.

Acknowledgments

Support for the present work was provided by the James S. McDonnell Foundation (M.B.) and the German Research Foundation (DFG), Emmy Noether fellowship TO 409/1-3 and SPP grant TO 409/7-1 (M.T.).

References

1 Niv, Y. (2009) Reinforcement learning in the brain. *J. Math. Psychol.* 53, 139–154
2 Balleine, B.W. and O'Doherty, J.P. (2010) Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* 35, 48–69
3 Silver, D. *et al.* (2012) Temporal-difference search in computer Go. *Mach. Learn.* 87, 183–219
4 Toussaint, M. and Storkey, A. (2006) Probabilistic inference for solving discrete and continuous state markov decision processes. In *ICML' 06 Proceedings of the 23rd International Conference on Machine Learning* (Cohen, W. and Moore, A., eds), pp. 945–952, ACM
5 Solway, A. and Botvinick, M.M. (2012) Goal directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychol. Rev.* 119, 120–154
6 Rawlik, K. *et al.* On stochastic optimal control and reinforcement learning by approximate inference. In *Proceedings, International Conference on Robotics Science and Systems (RSS 2012)* (Roy, N., ed), MIT Press (in press)
7 Toussaint, M. (2009) Robot trajectory optimization using approximate inference. In *Proceedings of the 26th International Conference on Machine Learning (ICML)* (Danyluk, A.P. *et al.*, eds), p. 132, ACM
8 Kappen, H.J. *et al.* (2012) Optimal control as a graphical model inference problem. *Mach. Learn.* 87, 159–182
9 Friston, K. *et al.* (2010) Action and behavior: a free-energy formulation. *Biol. Cybern.* 102, 227–260
10 Doya, K. *et al.*, eds (2006) *The Bayesian Brain: Probabilistic Approaches to Neural Coding*, MIT Press
11 Gold, J.I. and Shadlen, M.N. (2007) The neural basis of decision making. *Annu. Rev. Neurosci.* 30, 535–574
12 Sanborn, A.N. *et al.* (2010) Rational approximations to rational models: alternative algorithms for category learning. *Psychol. Rev.* 117, 1144–1167
13 Buesing, L. *et al.* (2011) Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comput. Biol.* 7, e1002211
14 Lau, B. and Glimcher, P.W. (2008) Value representations in the primate striatum during matching behavior. *Neuron* 58, 451–463
15 Padoa-Schioppa, C. and Assad, J.A. (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441, 223–226

1364-6613/$ – see front matter
http://dx.doi.org/10.1016/j.tics.2012.08.006 Trends in Cognitive Sciences, October 2012, Vol. 16, No. 10