Optimal planning to plan: People partially plan based on plan specificity

Mark K. Ho[1] (mho@princeton.edu), David Abel[2] (david_abel@brown.edu), Jonathan D. Cohen[1] (jdc@princeton.edu), Michael L. Littman[2] (mlittman@cs.brown.edu), Thomas L. Griffiths[1] (tomg@princeton.edu)

[1]Department of Psychology, Princeton University, Princeton, NJ, USA

[2]Department of Computer Science, Brown University, Providence, RI, USA

Abstract

Planning requires simulating future choices and consequences. This process is costly. But, it is also useful since it allows people to make choices *in the now* that have desirable future outcomes. What is a rational way to balance the immediate computational costs and future benefits of planning? Here, we argue that this involves *planning to plan*—adaptively deciding what actions to plan and when to plan those actions. To formalize this intuition, we develop the ideas of *partial planning* and *information-theoretic simulation costs*. Together, these allow us to define a novel Bellman objective that includes both environmental rewards and planning costs, which we solve using a gradient-based planning-to-plan algorithm. A key prediction of our account is that when the value of an immediate action depends on a *more specific* plan, the computational cost associated with that action will be higher. To test this qualitative prediction, we measure participant response times when solving a Gridworld task. We find evidence for our account of planning costs, indicating that people rationally plan to plan. Our formulation and results provide new insight into the meta-planning processes that support the scale and sophistication of human problem solving.

Keywords: Planning; Bounded Rationality; Simulation; Meta-Planning; Information Theory

Introduction

People's decisions are often informed by their long term goals. But determining how an action in the current moment relates to consequences in the future is computationally costly. It requires thinking not only about the current action, but also about future possible actions and contingencies. For example, imagine you are currently at home and want to get to the airport to catch a flight. How do you determine what to do right now? What would you think about right now? Clearly, some details are important to your current decision: When to call a car or whether there will be traffic is very relevant. But there are other details that, although not currently relevant, may become relevant in the future. For instance, after going through airport security, you will need to turn either left or right to get to the flight gate. The action you take then is important, but you do not have to commit to a specific course of action while at home. Rather, you can reasonably wait until you arrive at the airport before thinking through that aspect of your plan. Put another way, you could be strategic about what actions you plan and when you plan those actions. This is a form of meta-planning that we call "planning to plan".

Here, we describe a general formalism for planning to plan that incorporates several ideas. First, based on work on planning as probabilistic inference (Todorov, 2009; Botvinick & Toussaint, 2012), we develop the notion of a *partial plan*. Partial plans are representations of what an agent will do, but it allows for some actions to be specified in more detail than other actions. This captures the intuitive notion of thinking through certain specific decisions (e.g., when to call a car) while remaining non-committal on other decisions (e.g., whether to turn left or right after security). Second, we draw on ideas from *information theoretic bounded rationality* (Tishby & Polani, 2011; Rubin, Shamir, & Tishby, 2012; Ortega & Braun, 2013) to characterize the computational costs associated with planning. This allows us to express the cost of planning in terms of the minimum number of bits required to re-encode a new plan from a default plan. Finally, we treat the problem of planning to plan as a sequential decision-making problem in which an agent must partially plan at each timestep, taking into account task rewards, planning costs, and future opportunities to partially plan.

A key prediction of our account is that people will construct a partial plan that maximizes task rewards and minimizes immediate planning costs. In particular, if deciding what action to take requires more specific planning, then planning costs are higher. We test this prediction by having participants play a simple Gridworld navigation task and measuring the amount of time it takes them to take their first action as a function of the optimal partial plan from their starting state. Our results show that people are sensitive to this feature of a task, which supports our account of planning to plan.

Background

Planning and problem solving is often understood as a form of search over a problem representation (Newell & Simon, 1972). However, brute force search for an optimal plan is often computationally intractable, which is why computer scientists have developed a range of methods to make planning more managable. This includes methods such as heuristic search (Pearl, 1984) and hierarchical planning (Sacerdoti, 1974), among many others. Meanwhile, psychologists and neuroscientists have extensively documented the shortcuts and simplifications that people use to efficiently organize their thoughts and behaviors (Tversky & Kahneman, 1974; Lashley,

1951).

Given the need to simplify planning, one might ask: What if people strategically constructed plans that were simple enough to use in the moment but effective enough to make good long term decisions? Several lines of research in the existing literature provide insight into this question. For example, work on rational meta-reasoning in humans (Griffiths, Lieder, & Goodman, 2015) and anytime algorithms (Dean & Boddy, 1988) highlights how agents can make decisions while balancing the computational costs of those decisions. Additionally, work on intertemporal representation (Trope & Liberman, 2003) demonstrates how people's construal of different aspects of the world changes as a function of time and context. The work presented here attempts to develop a general formal framework that connects these ideas and test them in people.

Model

We first review the basic formalism for sequential decision-making and then describe our account of partial planning, information-theoretic planning costs, and a novel recursive Bellman objective.

Markov Decision Processes

The standard formalism for sequential decision-making is the Markov Decision Process (MDP) (Puterman, 1994). A ground MDP M is a tuple $\langle S, A, T, R, \gamma \rangle$, which consists of a set of states S; a set of actions A; a probabilistic transition function $T : S \times A \to \Delta(S)$; a reward function $R : S \times A \times S \to \mathbb{R}$; and a discount rate $\gamma \in [0,1)$ [1].

A policy, $\pi : S \to \Delta(A)$, maps states to distributions over actions, and the value function of a policy, $V^\pi : S \to \mathbb{R}$, is the expected discounted cumulative reward that an agent receives for following π from each state. The *optimal value function* is the unique fixed point of the Bellman equations (Bellman, 1957), for all $s \in S$:

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} T^a_{s,s'}[R^a_{s,s'} + \gamma V^*(s')]. \tag{1}$$

Additionally, we can define an optimal state-action value function $Q^*(s,a) = \sum_{s' \in S} T^a_{s,s'}[R^a_{s,s'} + \gamma V^*(s')]$, for all $s \in S, a \in A$. An *optimal policy* π^* is any policy π such that $V^\pi(s) = V^*(s)$ for all $s \in S$.

Partial Plans

We formalize partial plans in two steps. First, we define a "simulated" MDP, $\tilde{M}$, that corresponds to an agent's model of the actual task, M. We focus on the relationship between ground states S and simulated states $\tilde{S}$, and assume that $\tilde{M} = M$. This allows us to express an agent's plan from different states. For instance, the simulated action distribution at simulated state $\tilde{s}$ from ground state s is denoted $\tilde{\pi}(a \mid \tilde{s}; s)$.

Second, we allow an agent to control their partial policy at a state s via a *temperature assignment* over simulated states $\tilde{s}$. This captures the degree of optimal planning that an agent engages in at each simulated state. Formally, a temperature assignment from state s is $\beta(\cdot; s) : \tilde{S} \to \mathbb{R}_{\geq 0}$. An assignment defines soft-Bellman equations over simulated states:

$$\tilde{\pi}^\beta(a \mid \tilde{s}; s) \propto \exp\left\{ \tilde{Q}^\beta(\tilde{s}, a; s)\beta(\tilde{s}; s) \right\}, \tag{2}$$

$$\tilde{V}^\beta(\tilde{s}; s) = \sum_a \tilde{\pi}^\beta(a \mid \tilde{s}; s)\tilde{Q}^\beta(\tilde{s}, a; s), \tag{3}$$

$$\tilde{Q}^\beta(\tilde{s}, a; s) = \sum_{\tilde{s}'} T^a_{s,s'}\left[R^a_{s,s'} + \gamma \tilde{V}^\beta(\tilde{s}'; s) \right]. \tag{4}$$

Larger temperatures entail more optimal planning at a simulated state, and the interaction of temperatures over the entire model results in a partial plan. The temperature assignment controls how information about future rewards propagates through simulated states in a model.

Information Theoretic Planning Costs

Work on information theoretic bounded rationality provides inspiration for our formulation of planning costs (Tishby & Polani, 2011; Rubin et al., 2012; Ortega & Braun, 2013). In particular, we quantify the cost of a simulated partial plan $\tilde{\pi}$ in terms of the sum of Kullback-Leibler (KL) divergences (denoted D_{KL}) from a default policy $\bar{\pi}$ over all states:

$$C(\tilde{\pi}, \bar{\pi}) = \sum_{\tilde{s} \in \tilde{S}} D_{\mathrm{KL}}\left[\tilde{\pi}(\cdot \mid \tilde{s}) || \bar{\pi}(\cdot \mid \tilde{s}) \right]. \tag{5}$$

The KL-divergence is $D_{\mathrm{KL}}[p||q] = \sum_x p(x) \log(\frac{p(x)}{q(x)})$ for distributions p and q with the same support (Cover & Thomas, 1991). Following previous work (Gottwald & Braun, 2019), we set $\bar{\pi}$ to be the uniform distribution at all states.

Planning-to-Plan Bellman Objective

We can now define the problem of planning to plan by nesting partial planning inside a meta-planning problem that includes planning costs:

$$V^*_\lambda(s) = \max_{\beta(\cdot;s)} \left\{ \sum_{a \in A} \left[\tilde{\pi}^\beta(a \mid s; s) \sum_{s' \in S} T^a_{s,s'} \left[R^a_{s,s'} + \gamma V^*_\lambda(s') \right] \right] - \lambda C(\tilde{\pi}^\beta, \bar{\pi}) \right\}. \tag{6}$$

This modified Bellman objective extends the original Bellman equations (Equation 1) in two ways. First, the ground action distribution at the current state, $\tilde{\pi}^\beta(a \mid s; s)$, results from a planning process rather than being directly chosen. Second, planning costs are expressed through the term $\lambda C(\tilde{\pi}^\beta, \bar{\pi})$, where $\lambda \in \mathbb{R}$ is a planning cost weight. Equation 6 thus defines how an agent should partially plan their current decision given future partial planning.

[1] $\Delta(X)$ is the simplex over elements $x \in X$.

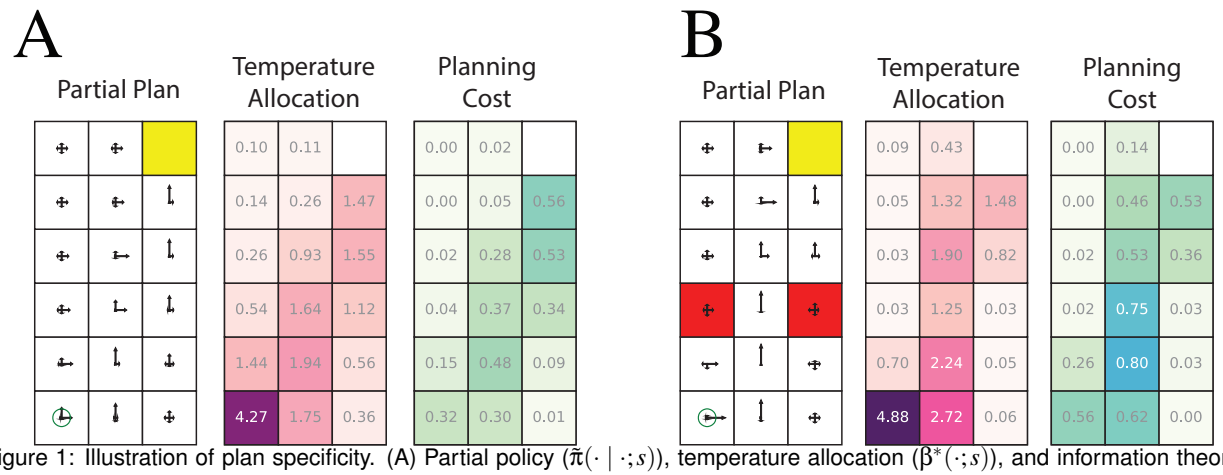

Figure 1: Illustration of plan specificity. (A) Partial policy ($\tilde{\pi}(\cdot \mid \cdot; s)$), temperature allocation ($\beta^*(\cdot; s)$), and information theoretic planning costs ($C(\tilde{\pi}, \bar{\pi})$) when at the state on the lower left (green circle) on a grid with no obstacles. (B) In the presence of obstacles, the decision at the lower left requires a more specific partial plan, which is more costly from an information theoretic planning perspective. Yellow tile is an absorbing state worth +10; red tiles are -5; step costs are -0.1; $\gamma = .95$; $\lambda = 0.01$.

Partial Plan Specificity and Planning Costs

In our formulation, an agent optimizing Equation 6 is minimizing planning costs, which means they will only simulate future decisions that are *relevant* to the current decision. Conversely, if making the current decision is contingent on a large number of simulated future decisions, then planning costs will be higher. Put simply, more specific planning involves greater planning costs.

To illustrate this property, consider the simulation results shown in Figure 1. We constructed a simple Gridworld in which an agent navigated from the lower left corner to an absorbing goal in the upper right. When there are obstacles, the agent needs to engage in more specific planning, which is more costly.

Experiment

To test whether people are sensitive to plan specificity as predicted by the model, we designed a simple Gridworld path-planning task in which the length of an optimal plan was held constant, but the required specificity of plan was manipulated. This was done by placing costly obstacles that blocked all but a few shortest path to a goal (Figure 2a). To assess how much initial planning people engaged in, we measured the amount of time before they took their first action. Our key qualitative prediction is that people will take longer to respond when they must identify a more specific plan to make their first action.

Materials and Procedure

Forty participants were recruited via Amazon Mechanical Turk to participate in our study using psiTurk (Gureckis et al., 2016). After familiarizing themselves with the mechanics of the task, each participant was given 48 versions of the grid in Figure 2a that varied by six initial start states and transformations based on the eight symmetries of a square (Figure 2b). The order of the rounds were one of eight pre-determined random sequences and was counterbalanced. The yellow goal state was worth 10 points and red squares were -5 points (5 points = 1¢ bonus). Each round began with a blank 9×9 grid. When participants pressed the space bar, the round was immediately loaded and actions could be taken in any of the four cardinal directions using the arrow keys. Initial state response times (RTs) were recorded by comparing the time between when the round loaded and the first action was taken.

Results

One participant was excluded from analysis due to missing data. To assess the influence of *goal distance* and *obstacles*, we ran a mixed-effects linear model with initial state log-transformed reaction times as the dependent variable. By-participant intercepts and round number slopes were included as random effects, and the initial state Manhattan distance to the goal, whether the initial state was on the side with obstacles, and their interaction were included as both random and fixed effects. This enables us to control for individual variance as well as learning effects across the task to determine the influence of plan specificity on reaction times. Figure 2c shows boxplots of log-transformed reaction time for the six different starting locations.

Consistent with our account's predictions, reaction times were faster when participants started on the side without obstacles (Fixed effect of obstacles: $\beta = 0.29, \text{SE} = 0.04, t(81.50) = 6.95, p < .001$). Moreover, not only were people faster the closer they were to the goal (Fixed effect of goal distance: $\beta = 0.05, \text{SE} = 0.004, t(39.53) = 11.20, p < .001$), but the effect of distance was stronger on the obstacle side (Fixed effect of interaction: $\beta = -0.05, \text{SE} = 0.005, t(60.60) = -10.97, p < .001$). This provides experimental evidence that people are sensitive to plan specificity, consistent with our model of planning to plan.

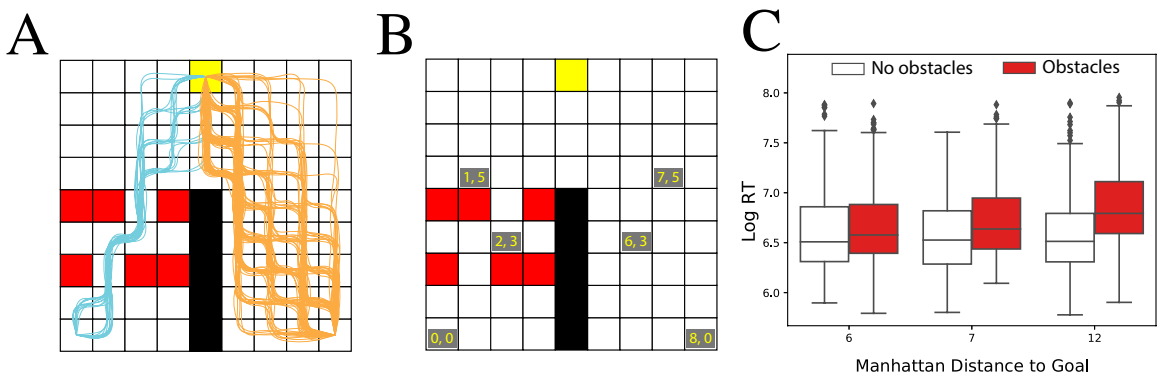

Figure 2: (A) Gridworld used in to test the effect of specificity on partial planning (Yellow tile is a +10 goal state, red tiles are -5). There are 2^{27} distinct optimal paths when starting on the lower-right corner of the grid (orange lines) and only 2^{7} when starting at the lower-left corner (blue lines) because of the red obstacles (random simulated trajectories shown). This means partial planning from states on the left will tend to be more computationally costly than ones on the right since it requires committing to a more specific path at the first timestep. (B) Yellow numbers indicate the six starting states tested in the experiment. (C) Box-plots of log reaction times on initial state by Manhattan distance to goal and whether the trials were on the side of obstacles.

Discussion

Planning is useful because it allows agents to make decisions informed by future possibilities. But determining how future actions and consequences influence current decisions is computationally costly. We argue that agents must then be strategic about what they plan and when they plan: Resource limited agents should plan to plan. Here, we have formalized the notion of planning to plan by using partial planning and information theoretic costs to define a novel Bellman equation. Our account predicts that when a good decision requires a more specific plan, planning costs will be higher. We test this qualitative prediction in people by measuring reaction times when navigating Gridworlds with and without obstacles in the way and find support for this prediction. This provides evidence that people plan to plan.

References

Bellman, R. (1957). *Dynamic programming*. Princeton University Press.

Botvinick, M. M., & Toussaint, M. (2012). Planning as inference. *Trends in cognitive sciences*, *16*(10), 485–488.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. New York, USA: John Wiley & Sons, Inc. doi: 10.1002/0471200611

Dean, T. L., & Boddy, M. S. (1988). An analysis of time-dependent planning. In *Aaai* (Vol. 88, pp. 49–54).

Gottwald, S., & Braun, D. A. (2019). Bounded rational decision-making from elementary computations that reduce uncertainty. *Entropy*, *21*(4).

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. doi: 10.1111/tops.12142

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Lashley, K. S. (1951). The problem of serial order in behavior. In *Cerebral mechanisms in behavior; the hixon symposium.* (pp. 112–146). Oxford, England: Wiley.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall, Englewood Cliffs, NJ.

Ortega, P. A., & Braun, D. A. (2013, mar). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *469*(2153), 20120683–20120683. doi: 10.1098/rspa.2012.0683

Pearl, J. (1984). *Heuristics: Intelligent search strategies for computer problem solving*. Addison-Wesley.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc.

Rubin, J., Shamir, O., & Tishby, N. (2012). Trading Value and Information in MDPs. In (pp. 57–74). Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-24647-0_3

Sacerdoti, E. D. (1974). Planning in a hierarchy of abstraction spaces. *Artificial Intelligence*, *5*(2), 115 - 135. doi: https://doi.org/10.1016/0004-3702(74)90026-5

Tishby, N., & Polani, D. (2011). Information Theory of Decisions and Actions. In *Perception-action cycle* (pp. 601–636). New York, NY: Springer New York. doi: 10.1007/978-1-4419-1452-1_19

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, *106*(28), 11478–11483. doi: 10.1073/pnas.0710743106

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological review*, *110*(3), 403.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. doi: 10.1126/science.185.4157.1124