Information sampling for contingency planning

Ili Ma (i.ma@fsw.leidenuniv.nl)
Department of Developmental and Educational Psychology, Leiden University, Leiden, the Netherlands

Wei Ji Ma (weijima@nyu.edu)
Department of Psychology and Neural Science, New York University, NY, USA

Todd Gureckis (todd.gureckis@nyu.edu)
Department of Psychology, New York University, NY, USA

Abstract

From navigation in unfamiliar environments to career planning, people typically first sample information before committing to a plan. However, most studies find that people adopt myopic strategies when sampling information. Here we challenge those findings by investigating whether contingency planning is a driver of information sampling. To this aim, we developed a novel navigation task that is a shortest path finding problem under uncertainty of bridge closures. Participants (n = 109) were allowed to sample information on bridge statuses prior to committing to a path. We developed a computational model in which the agent samples information based on the cost of switching to a contingency plan. We find that this model fits human behavior well and is qualitatively similar to the approximated optimal solution. Together, this suggests that humans use contingency planning as a driver of information sampling.

Keywords: Planning; Uncertainty; Decision-Making; Markov Decision Process

Introduction

Acquiring information for the purpose of planning is common in the real world; ranging from navigation to searching for information to plan one's career. Such information acquisition typically serves to avoid large costs, such as avoiding a time-consuming traffic jam, or choosing an education that does not prepare for a desired profession. In this study, we examine how contingency planning drives information sampling. We operationalize planning as the mental simulation of potential futures, and consider uncertainty that arises from incomplete knowledge of the environment. We define a contingency plan as an alternative when the original plan turns out to be infeasible or too costly. For example, in navigation, taking a detour after encountering a roadblock can be seen as a contingency plan. Importantly, the cost of switching to a contingency plan can be high, e.g., having to turn back and drive far back to access an alternative route. The basic hypothesis of the current study is that people efficiently collect information to reduce the costs of switching plans.

If sampling information for the purpose of planning is indeed governed by the cost of switching to the contingency plan, then sampling information for planning is a form of 'interested active sampling' (Chambon et al., 2020; Gureckis & Markant, 2012; Markant & Gureckis, 2014; Rouault et al., 2021). Interested sampling specifies that the sampling decisions are guided by the economic utility of the information. In contrast, disinterested sampling specifies that agents sample to reduce uncertainty about the world (i.e., reduce entropy) without considering the economic utility of the sample. The distinction between "interested" and "disinterested" sampling has been an open question in the literature on information sampling for several decades (Chater, Crocker, & Pickering, 1998; Gureckis & Markant, 2012; Coenen, Nelson, & Gureckis, 2019). The present work adds to this question by assessing, specifically, if people value information to avoid costly replanning.

Intuitively, sampling information to avoid costly replanning seems to be a natural part of intelligent human behavior. For example, when travelling in an unfamiliar area, people gather information to determine how to reach their destination rather than taking a trial-and-error approach. Despite its ubiquity, planning paradigms within cognitive science (e.g. those described in (van Opheusden & Ma, 2019)) do not typically allow for sampling along a future plan. Decision making is usually about which next step to take, whereas in many everyday tasks we can consider a plan in advance and take actions to reduce uncertainty about the possible roadblocks in that plan.

We study a particular class of planning under uncertainty tasks called the Canadian Traveler problem (Papadimitriou & Yannakakis, 1991). Specifically, the task is to reach a goal while minimizing travel distance in a known graph but with parts of the roads (bridges) that might be blocked and therefore untraversable (by avalanches caused by heavy snowfall). Bnaya, Felner and Shimony (2009) developed a generalization of this paradigm that allows for sampling information, with the goal to minimize the sum of travel cost and sampling actions. In the current study, we take this approach to examine information sampling for the purpose of planning. The task is realistic, intuitive, and complex while still tractable to computational analysis.

Methods

Participants and procedure We collected data online from 109 participants via Amazon MTurk. After signing the consent form, participants completed our task, followed by the Barratt Impulsiveness Scale-11 (BIS-11; Patton et al., 1995), which measures trait impulsivity and the Future Orientation Scale, which measures the degree to which individuals perceive, anticipate, and plan for the future in daily life (Steinberg et al., 2009). These two questionnaires were included to assess whether the information sampling choices in our task correlated with planning in daily life. Since planning requires reasoning about future states, planning was expected to be inversely correlated with trait impulsivity.

Sheep Traveler Task is based on the structure of the original Canadian Traveler Problem (Papadimitriou & Yannakakis, 1991; Bnaya et al., 2009). We programmed the task in C#, using the Unity game engine for visualization. For each puzzle, participants viewed a map of bridges consisting of 10-60 bridges, their avatar's position (Figure 1, sheep), and a goal (Figure 1, flag). They were asked to navigate the avatar to the goal in as few actions as possible. Crucially, the status of a bridge could be "open", "closed", or "unknown". All bridges were "unknown" at the start of a new puzzle. The task consisted of two stages, in stage 1, the participant could sample the status of up to ten bridges anywhere in the puzzle. Once the participant had either sampled the status of ten bridges or clicked a button to indicate being finished with sampling, they enter stage 2. In stage 2, the participant had to get the sheep to the goal by moving it across bridges. For unknown bridges, the status (open or closed) was revealed when the participant tried to move the sheep over it. An action in stage 1 was sampling a bridge's status, in stage 2 an action was a move or attempt to move the sheep across a bridge. Each action incurred a cost of \$0.05. The task was completed after 85 puzzles, then 3 puzzles were pseudo-randomly chosen and their summed cost was subtracted from a \$8.50 participation fee. Participants were told that some bridges cause a long detour when closed, so they could keep the cost lower by sampling wisely in stage 1. Participants were informed of the probability that a bridge was closed was 0.2 and that all closures were independent of each other. They completed 10 practice puzzles to familiarize themselves with the task and the probability of closed bridges. This was followed by a quiz to check the comprehension of the instructions before starting the task. An incorrect response on the quiz led the participant back to the tutorial. Participation was terminated if there were more than 3 incorrect attempts. Otherwise, the main task with 85 puzzles started. The order of the puzzles was randomized per participant.

The puzzle finished when: a) the sheep avatar reached the target; b) when the bridges of known status showed with certainty that the puzzle had no solution (i.e., the sheep cannot reach the goal). The puzzles (bridges, the sheep's starting position, goal location, and a hidden bridge status map) were designed to distinguish between the (near-)optimal solution and a myopic policy. Specifically, puzzles were designed such that a myopic policy would result in less reward.

Computational model The task can be described as a Markov Decision Process with a large state space. In stage 1, actions, a, are remote samples. In stage 2, actions are moves to an adjacent node and attempts to do so. The state, s, is defined by the combination of the avatar's location and board state. The board state is the collection of locations and statuses of all bridges which can be unknown, open, or closed.

Our main model accounts for actions in both the remote sampling stage and the moving stage. The intuition for the model is that people think about the possible paths that they could choose from in the moving stage. On those paths they evaluate how long the detour would become if a given bridge is closed. This allows the model to identify *bottlenecks*, which are bridges that would cause a long detour when closed. These two aspects of the model are then combined such that bridges on paths with a shorter effective length and long potential detour are more likely to be sampled in stage 1. These evaluations and sampling continue until the potential detour is shorter than a free parameter (a stopping threshold) or ten samples, which is the maximum samples per puzzle. In stage 2, the agent chooses a path proportional to the effective length.

Formally, we define a path, P, as a contiguous set of non-closed bridges without repetitions (without moving back and forth), leading from the avatar's current location to the goal location. The *state* of a bridge on a traversable path is open or unknown. We denote the set of the states of all the bridges along a path P by S_P. We define the *effective length* L of a path as

$$L(P,S_P) \equiv N_{\text{open}}(P) + \omega N_{\text{unknown}}(P), \tag{1}$$

where N_{open} and N_{unknown} are the numbers of open and unknown bridges in the path, respectively, and the number of unknown bridges on the path is weighted by an uncertainty aversion parameter ω. The uncertainty aversion heuristic is based on well-established findings in the field of risky decision-making, showing that people prefer to choose options with a known probability of winning over an unknown probability of winning, known as the Ellsberg paradox (Ellsberg, 1961).

The value of a path is its negative effective length, biased by a preference k for paths on which bridges were previously sampled:

$$V_{\text{path}}(P,S_P) = -L(P,S_P) + k\chi_P, \tag{2}$$

where $\chi_P = 1$ for all paths P that contain any previously sampled bridge and 0 otherwise.

We assume that the probability of choosing path P_i is given by a softmax on this path value:

$$p(\text{choose } P_i) = \frac{e^{\beta_{\text{path}} V_{\text{path}}(P_i, S_{P_i})}}{\sum_j e^{\beta_{\text{path}} V_{\text{path}}(P_j, S_{P_j})}}. \tag{3}$$

Note that the participants do not explicitly have to indicate a path but rather, they choose bridges to sample. It is assumed that these bridges are on paths that they consider for moving the sheep to the goal.

We now consider what happens when sampling causes the status of an unknown bridge, indexed by j, on a path P to

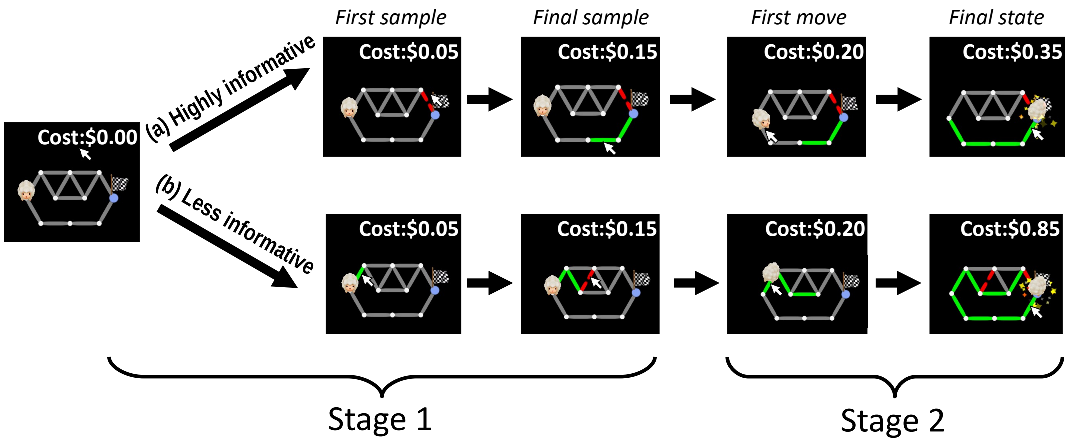

Figure 1: Example of sampling strategies in a small puzzle in the remote sampling in the Canadian Travel Task. The upper row (a) shows an agent who selects a highly informative sample, as discovering that this bridge is closed saves the agent from having to take a long detour. The lower row (b) shows an agent who selects a less informative sample, as this sample does not save the agent from a long detour in stage 2, even if it were closed. The sampling strategies influence the reward, as agent "a" ends the puzzle with a lower total cost than agent "b". In the actual task, the puzzles consisted of more bridges than in this example.

change to open or closed. If it were to change to open, the path P would remain valid but the status S_P would change to $S'_{P,j}$. If it were to change to closed, then the agent would have to consider an alternative path P'_j, which we will refer to as a *detour*. This detour length is given by the path length to bridge j plus the shortest path from bridge j to the goal under the assumption that all unknown bridges are open. We then calculate the cost incurred by taking the detour as the difference between the effective length of the detour if the jth bridge were closed and the effective length of the original path if that bridge were open:

$$\Delta L_j(P) = \underbrace{L(P'_j, S_{P'_j})}_{\text{length of detour}} - \underbrace{L(P, S'_{P,j})}_{\text{length of primary path}} . \tag{4}$$

(In the special case that a closed bridge results in an unsolvable puzzle, $\Delta L_j(P)$ is the length from the agent to that bridge.) To compute the expected detour cost of bridge j, we then consider all paths P_i that run through j, weighting them by the probability that the path on which the bridge lies is actually chosen:

$$E\left[\Delta L_j\right] = \sum_i \phi_{ij} p(\text{choose } P_i)\Delta L_j(P_i), \tag{5}$$

where $\phi_{ij} = 1$ if path P_i passes through bridge j and $\phi = 0$ otherwise. We now define a value function for bridges. The value of sampling bridge j is

$$V_j = E\left[\Delta L_j\right] - \lambda d_j, \tag{6}$$

where d_j reflects the (shortest) distance from bridge j to the position of the avatar, measured along open and unknown bridges. This "near avatar bonus" accounts for the fact that some people might have a preference for sampling closer to the avatar than on other bridge locations along a path. Indeed, people sometimes adopt myopic strategies when sampling information (Schulz & Gershman, 2019). Such myopic behavior is usually not optimal in our task because it does not save you from a long detour.

We assume that the probability to stop sampling before all ten samples are used is a softmax function of the highest unknown bridge value:

$$p(a = \text{stop}|S) = \left(1 + e^{-\beta_{\text{stop},0}+\beta_{\text{stop},1}\max_j V_j}\right)^{-1} . \tag{7}$$

Where $\beta_{\text{stop},0}$ is the stopping intercept and $\beta_{\text{stop},1}$ is the softmax temperature. Finally, if the agent does sample, we assume that the sampling probabilities are given by a softmax on the bridge values:

$$p(a = \text{sample bridge } j|S, a \neq \text{stop}) \propto e^{\beta_{\text{bridge}}V_j} . \tag{8}$$

The full model has seven free parameters: ω, k, d_j, β_{path}, β_{bridge}, $\beta_{\text{stop},0}$, and $\beta_{\text{stop, 1}}$. The model was fitted with the fmincon function in Matlab using 100 random initializations to avoid local minima.

Alternative detour models We fitted the full model described above and compared this to all versions without the three heuristic free parameters. Specifically, we compared the full model to the model without: 1. uncertainty aversion ω, 2. the preference to keep sampling on the same path k, 3. the near avatar bonus d_j, 4. uncertainty aversion and the near avatar bonus, 5, uncertainty aversion and the same path preference, 6. same path preference and near avatar bonus, 7. uncertainty aversion, same path preference, and near avatar bonus. For model comparison, we computed the 95% confidence interval of the median difference in BIC between the full model and each lesion model. We used bootstrapping implemented in the R package "boot". We consider the difference not significant if the interval contains 0.

Alternative heuristic models One alternative hypothesis is that sampling is not driven by contingency plans, but rather by the features of the bridge network. To test this, we developed a heuristic Feature model, in which the agent's samples are based on weighted graph theoretic features. Specifically, for each bridge and on each observed state, we computed the following four features: 1. *Betweenness centrality*, which is the number of shortest paths from the sheep to the goal that run through the bridge, 2. *Distance to avatar*, which is the number of bridges between the current bridge and the avatar using the shortest not closed path, 3. *Detour length*, which is the shortest detour length from the current bridge to the goal if the current bridge were closed, 4. *Degree*, which is the number of connected bridges to the nodes attached to the current bridge. The model uses a weighted sum of these features, augmented with a constant term to determine V_j, the value of bridge j. The probability of stopping is given by Eq. (7), and the probability of sampling is given by:

$$p(a = \text{sample bridge } j | S, a \neq \text{stop}) \propto e^{V_j}. \qquad (9)$$

For this model, we also fitted the full model and compared this to all versions with dropped-out features using the 95% CI of the median BIC difference.

Approximating optimal behavior

Due to the large state space, computing the optimal solution through dynamic programming is near-intractable. We therefore approximate the optimal policy using Monte Carlo Tree Search (MCTS) and use its performance as a benchmark for human behavior (Silver et al., 2016). MCTS is a decision tree search algorithm in which evaluation of a node is done by simulating actions until the game ends, the reward is then determined and backpropagated to its parent nodes. In standard MCTS, the tree search is performed by iteratively building a tree where the nodes represent states and the edges that connect the nodes represent actions. However, in our task, an action did not always lead to the same state. Stochasticity needed to be introduced to the tree because actions in which the agent sampled or attempted to cross an unknown bridge could result in either a state in which the bridge was open, or a state in which it was closed. Therefore, we used MCTS for stochastic environments, in which the tree had alternating action nodes and chance nodes (Veness, Ng, Hutter, Uther, & Silver, 2011). Action nodes describe the action space (i.e., all legal moves in the current state), and can these can have two child nodes, called chance nodes, which represent the resulting state (i.e., one node for an open bridge and one for a closed bridge). A single iteration can be described in four steps:

1. Node selection. A node is chosen for expansion
2. Expansion. The node is expanded by simulation
3. Rollout. A game is simulated until an end state is reached (either the sheep reached the goal or when it is proven that the sheep cannot reach the goal). We applied greedy rollouts with exploration noise to avoid bias in the trees.
4. Backpropagation. The reward is backpropagated up the tree to update the value of the parent nodes. The reward is the negative number of actions in stage 1 and 2 combined, as the task goal was to minimize the total number of actions. The visit count of the parent nodes is also updated by 1.

We ran the algorithm for minimally 100,000 iterations per node, calibrated to ensure that a stable policy is reached. The Upper Confidence Bound for Trees (UCT) was used to choose the node to expand (Kocsis & Szepesvári, 2006). UCT balances between exploitation of promising nodes and exploration of less often visited nodes. The leaf node with the highest UCT value was selected for expansion. The UCT formula is as follows: $\text{UCT} = \frac{-L_i}{N_i} + c\sqrt{\frac{\ln(N_i)}{n_i}}$, where L_i is the cumulative number of actions of all games played at node i to complete the puzzle. n_i is the number of visits to node i, N_i is the total number of visits to the root node, and c is the exploration weight, here set to 25 (corresponding to the theoretically optimal weight of $\sqrt{2}$ expected actions) to avoid bias in the tree.

The resulting MCTS policy is not fully deterministic, for example when a state contains action nodes in which visit counts are uniformly distributed. We therefore generated data of 100 MCTS agents to compare their policy to human data.

Results

How much information did people sample? We were first interested in how often people sampled information overall. In stage 1, people sampled before moving the sheep in stage 2 (m = 7.07; SD = 2.94 out of the ten possible remote samples per puzzle). Next, we approximated the optimal policy using MCTS and compared sampling quantities between MCTS and humans. Using a Mann-Whitney U test, we found no significant difference in the number of samples between humans and MCTS in stage 1 of the task (U = 814, p = 0.491). However, human participants did make more moves in stage 2 than was approximately optimal according to MCTS (U = 440, p = 0.002, Figure 2a) and overall took more actions to complete the task than was optimal (U = 551, p = 0.02). Interestingly, this suggests that humans were roughly well calibrated to the optimal amount of information sampling but possibly the specifics choices for these samples were not as effective at improving the avatars movements. In addition, after sampling a closed bridge, people often switched to sampling on the new shortest path (Figure 2b). This suggests that people switched to sampling on the next best plan after discovering that the original plan could not lead to the goal.

Detour model shows the best fit The full model without a preference k to sample bridges on the same path fitted best.

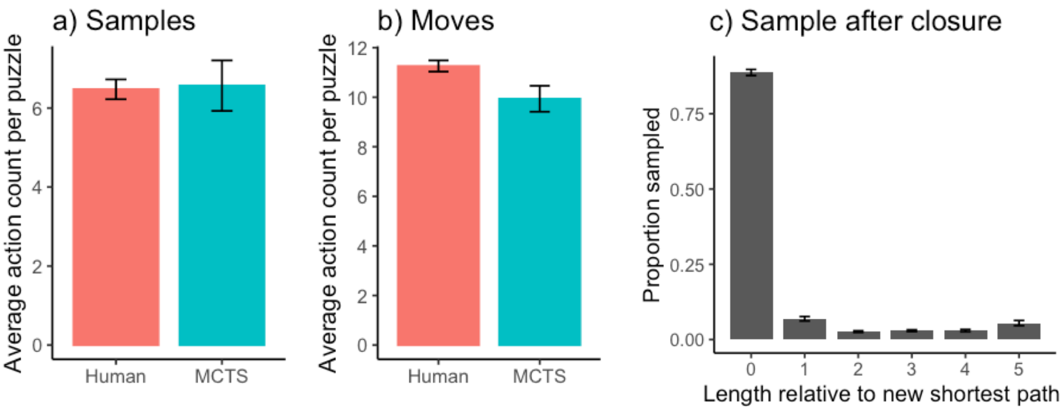

Figure 2: Sampling data. A. Average number of actions in each stage for humans (orange) and the optimal policy approximated using Monte Carlo Tree Search (blue). "Samples" are the remote samples in stage 1. "Moves" are the sheep avatar's moves in stage 2. The error bars show the between-subjects standard error of the mean. B. The proportion of samples on the new shortest path after sampling a closed bridge (human data). On the x-axis, 0 = the new shortest path, 1 = a path that is one bridge longer than the new shortest path, 2= two bridges longer, etc.

This model included free parameters for uncertainty aversion and a near avatar bonus (Figure 3). This model also fitted somewhat better than the best Feature model (median BIC difference = -213, 95% CI [-355,-9]). Thereby suggesting that sampling is driven by contingency planning, rather than the visual representation of the graph's features.

Action-by-action fits The primary choice data for this task are rich and high dimensional (structured on a graph and depending on the agent's position). As a result, the match between the model's simulated data and the human data is best evaluated on key features of the behavior. To visualize the relationship between data and model fit, we plotted the probability of sampling an edge as function of various graph theory features that best describe the relation of each bridge to the full puzzle. We also generated data using the parameter estimates of our computational model to compare this to human data. Figure 3 shows the human data and the data generated by the fitted model as function of each graph feature. This figure shows that the model fitted well and accounted for qualitatively similar patterns as the human data. As expected, people preferred to sample bridges that would cause a long detour length if closed (high detour length), were on the shortest paths (high betweenness centrality), and had few other bridges connected to them (low degree). The intuition for preferring bridges that are connected to nodes with a low degree is that alternative paths are less easily accessible from those nodes. In an extreme case the bridge can be a bottleneck. As shown by the U-shaped curve for distance to avatar, the near avatar bonus induced a bias to sample bridges that were close to the goal, and in some cases biased to sampling bridges that were close to the avatar (Figure 3b). The computational model also fitted human data better than the optimal approximation derived by MCTS Figure 3). The fitted model also performed better than a random strategy (median BIC difference = -126 in favor of our main model, with 95% CI [-171, -55]).

Individual differences We next characterized individual differences in sampling strategies. By using Ward's hierarchical clustering method on the Euclidean distance between estimates in the 6D parameter space, we found similarities between subjects that translate into distinct behavioral strategies. Specifically, we found three distinct sampling strategies; sparse sampling near the goal, sampling from the goal towards avatar on the shortest path, sampling from the avatar towards the goal on the shortest path. Out of these three strategies, sparse sampling near goal was the least prevalent (about 10% of the subjects), while the other two strategies were almost evenly prevalent. In our puzzles, sampling from the goal towards the avatar is usually the better strategy as it can save the agent from having to take a long detour. Sampling from the avatar towards the goal, however, might reflect a myopic strategy.

We found a modest negative correlation between scores on the Future Orientation Scale and the near avatar bonus parameter (r = -0.275, p = 0.006). This shows that people who are less future oriented in their daily lives also sampled more myopically on the task. There were no significant correlations between uncertainty aversion estimates and the scores on the Future Orientation Scale or between trait impulsivity as measured with the BIS-11 and either of the two parameter estimates (all $p \geq 0.18$).

Discussion

We aimed to identify guiding principles by which people sample for the purpose of contingency planning. Past studies have established that people rarely consider a longer time horizon when determining the sequence of information sampling (Meder, Nelson, Jones, & Ruggeri, 2019; Ma, Sanfey,

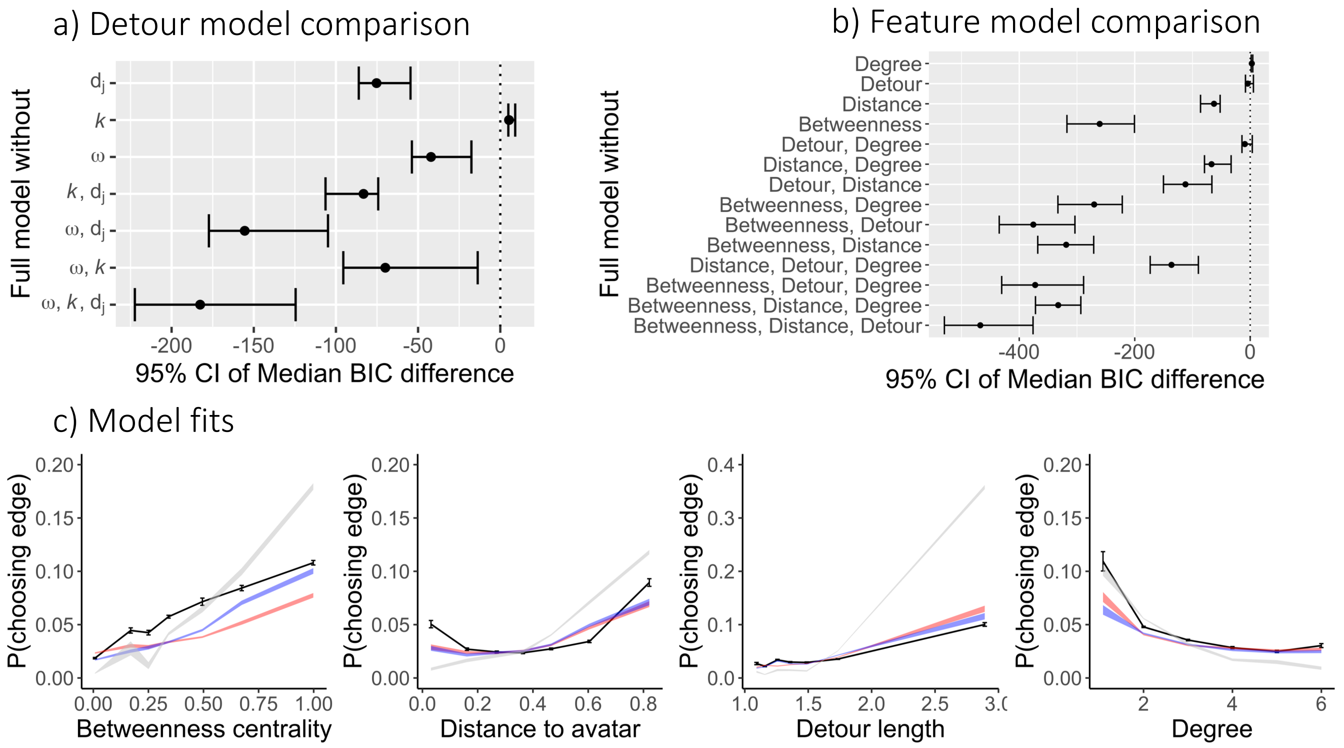

Figure 3: Model comparisons and model fits. A. Detour model comparisons. B. Feature model comparisons. The x-axis shows the 95% CI of the median BIC difference, which was computed using bootstrapping. A larger BIC difference indicates a better fit. The difference is not significant if the 95% CI contains zero. A. model comparisons for the Detour model. B. Model comparisons for the Feature model. C. Data and model fit as function of the four graph features. The black line graph with error bars show the man and SEM of the human data using quantile bins. The shaded area depicts the SEM of the generated data using the fitted model using the same bin centers as the human data. Red shaded region = Detour model. Blue = Feature model. Grey = the optimal policy approximated using MCTS. Note that the y-axis range of the detour length plot is larger.

& Ma, 2020). Instead of planning ahead for their information search they tend to adopt a myopic strategy, which is suboptimal (Nelson, Meder, & Jones, 2018). Yet, information sampling is often instrumental to planning and planning requires thinking about future states. In daily life, people often gather information to determine if their plans may or may not come to fruition. We designed a novel planning task that allows for information sampling prior to committing to a plan. In this task, the utility of a sample can be determined by thinking about one's future states in a given plan and compute the cost of switching to a contingency plan. The utility of the sample then increases with the cost of switching to a contingency plan. Using this task and a computational model of contingency planning, we find that people do in fact identify information that is relevant over a longer time horizon when they sample for the purpose of planning. This concept is somewhat distinct from the question if people use stepwise or optimal information utilities to guide search but does show how representations of future actions can structure information sampling.

The samples generated by our computational model of contingency planning were qualitatively similar to the approximated optimal solution derived with MCTS, and far outperformed a random strategy. We found that the computational model fitted human behavior reasonably well and better than a model in which the agent samples based on graph theoretic features, suggesting that people use the cost of switching to contingency plans to guide their sampling decisions. We added two heuristics to the final model; uncertainty aversion and a near avatar bonus. These improved the model fit, showing that people tend to avoid paths with a higher uncertainty and sometimes look insufficiently ahead. Insufficiently looking ahead was ostensibly sensitive to individual differences and related to diminished orientation to the future in daily life. Importantly, adding these heuristics does not abolish the effect of detour length, demonstrating that people use the costs of contingency plans to determine where to sample. A third heuristic, the bias to keep sampling on paths on which previous samples were drawn, did not improve the model fit and was therefore not included in the final model.

A recent study proposed that active information seeking reveals which plans an individual considers, as information search is directly observable whereas planning itself is not (Callaway et al., 2021). This paper used a navigation under uncertainty task in which information about rewards and losses associated with road sections could be sampled. In contrast to our work, they found that sampling was strongly influenced by a myopic bias similar to our "near avatar bonus". This discrepancy might be due to the fact that we intentionally designed our task to test whether contingency planning drives sampling. This resulted in puzzles in which some paths had a much higher cost of switching to the contingency plan (i.e., longer potential detour) than others.

One critical element of our model is the identification of the shortest paths and identifying the approximate length of the detour. It is likely that these computations are implemented efficiently by the perceptual system. Earlier work on human performance on the Traveling Salesman problem shows that humans far outperform simple construction algorithms, for review see (MacGregor & Chu, 2011). People also perform much better in terms of solution speed and accuracy when the Traveling Salesman problem is visually presented than when it is presented as a table with intercity distances (Polivanova, 1974). Nevertheless, a resource rational model of sampling for contingency plans might be a more realistic account. In future work this can for example be achieved by adding a pruning parameter to the model (Huys et al., 2012) such that the agent considers only a limited number of paths.

Information sampling studies typically show that people adopt myopic strategies when sampling information. Here we show that information sampling and planning are tied together and that people do consider their future states when sampling for the purpose of planning. Our work thereby connects the planning and sampling literature.

Acknowledgments

This work was supported by NSF Award #2021060 and NWO Rubicon award 019.173SG.015.

References

Bnaya, Z., Felner, A., & Shimony, S. E. (2009). Canadian traveler problem with remote sensing. In *Twenty-First International Joint Conference on Artificial Intelligence.*

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P., Lieder, F., & Griffiths, T. (2021). Human planning as optimal information seeking. *Manuscript in preparation.*

Chambon, V., Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., & Palminteri, S. (2020). Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour*, *4*(10), 1067–1079.

Chater, N., Crocker, M. J., & Pickering, M. J. (1998). The rational analysis of inquiry: The case of parsing. Oxford University Press.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2019). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, *26*(5), 1548–1587.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 643–669.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, *8*(3), e1002410.

Kocsis, L., & Szepesvári, C. (2006). Bandit based monte-carlo planning. In *European conference on machine learning* (pp. 282–293).

Ma, I., Sanfey, A., & Ma, W. (2020). The social cost of gathering information for trust decisions. *Scientific reports*, *10*(1), 1–9.

MacGregor, J. N., & Chu, Y. (2011). Human performance on the traveling salesman and related problems: A review. *The Journal of Problem Solving*, *3*(2), 2.

Markant, D., & Gureckis, T. (2014). Is it better to select or to receive? learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*, 94–122.

Meder, B., Nelson, J. D., Jones, M., & Ruggeri, A. (2019). Stepwise versus globally optimal search in children and adults. *Cognition*, *191*, 103965.

Nelson, J. D., Meder, B., & Jones, M. (2018). Towards a theory of heuristic and optimal planning for sequential information search.

Papadimitriou, C. H., & Yannakakis, M. (1991). Shortest paths without a map. *Theoretical Computer Science*, *84*(1), 127–150.

Polivanova, N. (1974). Some functional and structural features of visual intuitive components of a problem-solving process. *Voprosy Psikhologii*(4), 41–51.

Rouault, M., Weiss, A., Lee, J. K., Bouté, J., Drugowitsch, J., Chambon, V., & Wyart, V. (2021). Specific cognitive signatures of information seeking in controllable environments. *bioRxiv*, 2021–01.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology*, *55*, 7–14.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

van Opheusden, B., & Ma, W. J. (2019). Tasks for aligning human and machine planning. *Current Opinion in Behavioral Sciences*, *29*, 127–133.

Veness, J., Ng, K. S., Hutter, M., Uther, W., & Silver, D. (2011). A monte-carlo aixi approximation. *Journal of Artificial Intelligence Research*, *40*, 95–142.