Arbitrating between planning and habit in naturalistic environments

Ugurcan Mugan (umugan@u.northwestern.edu)
Northwestern University, Neuroscience and Robotics Laboratory, 2145 Sheridan Road
Evanston, IL 60208 USA

Malcolm A. MacIver (maciver@northwestern.edu)
Northwestern University, Neuroscience and Robotics Laboratory, 2145 Sheridan Road
Evanston, IL 60208 USA

Abstract

Research into the neural basis of decision making suggests that animals engage in habit- or plan-based action selection. The existence of these two modes of action selection raises an additional problem: how should animals arbitrate between them to efficiently allocate their time and computational resources? Here, we use a naturalistic task, spatial planning of a prey evading a predator in environments with varying complexity, to investigate how the arbitration between these decision making systems should be done. We identify a key signature of complex environments where planning becomes imperative—transitions between poorly and highly connected regions. We suggest an efficient approach, based on environmental connectivity, that switches between plan- and habit-based control during a task. This approach provides a unifying account of experimental data that shows vicarious trial and error at high cost choice points, as well as increased theta coherence between the hippocampus and the prefrontal cortex at transitions from closed to open regions—both situations where there is a transition in spatial connectivity.

Keywords: planning; habit; Markov decision processes; predator-prey interactions

Introduction

Studies of animal decision making reveal two distinct systems: habit- and plan-based action selection (Daw, Niv, & Dayan, 2005; Keramati, Dezfouli, & Piray, 2011). These two systems have primarily been associated with the lateral striatum and its dopaminergic afferents for habit (Yin & Knowlton, 2006), and the interaction between the hippocampus and the prefrontal cortex (PFC) for planning (Ito, Zhang, Witter, Moser, & Moser, 2015; Schmidt, Duin, & Redish, 2019). Prior research into the neural basis of planning in mammals has been related to the phenomena of nonlocal spatial representation in hippocampal activity through vicarious trial and error (VTE) (Johnson & Redish, 2007), and replay (Pfeiffer & Foster, 2013). While replay happens more frequently at the start of a trial and at the reward port (Mattar & Daw, 2018), VTE, which coincides with pronounced theta and gamma rhythms, has been shown to occur at high cost choice points. It has been suggested that an increase in theta coherence between the hippocampus and PFC is needed to sort through options as the hippocampus imagines potential outcomes (Benchenane et al., 2010). Interestingly, the coherence between the hippocampus and the PFC increases as rodents transition from a closed arm to an open area (Adhikari, Topiwala, & Gordon, 2010). Both the phenomena of VTE, and the modulation of theta coherence within trials, suggests that animals switch back and forth from habit- to plan-based action selection continuously.

Here, we propose a formalization that identifies high cost choice points, and accordingly arbitrates between habit- and plan-based action selection in dynamic and complex environments that model predator-prey interactions.

Methods

In both habit- and plan-based action selection, action choices are dependent on the current state of the animal (location of the prey and the predator). The prey's knowledge about the current state allows it to associate a given outcome with an action, enabling it to predict long-term reward. The habit solution to the problem of long-term reward prediction ("model-free" (Schultz, Dayan, & Montague, 1997)) simply assigns a value to an an action (or an action sequence) based on prior experience. Conversely, the planning solution ("model-based" (Dolan & Dayan, 2013)) relies on an 'action-outcome' knowledge structure to generate action sequences by simulating future states and their expected outcomes.

To study the arbitration of habit- and plan-based action selection in a dynamic, ethologically relevant, and naturalistic spatial navigation task, we used the formalization of reinforcement learning theory (Sutton & Barto, 2018). We created a survival task by implementing a simulated prey and predator both acting in a 15×15 gridworld. The aim of the prey was to get to a safety location (analogous to a burrow) while being aggressively pursued by a predator, which is on average $1.5\times$ faster (Elliott, Cowan, & Holling, 1977; Hedenström & Rosén, 2001). The prey was configured to have habit-based action selection and plan-based action selection with a preset number of states that it could forward simulate (5000). We generated naturalistic environments by adding randomly generated obstacles until a predetermined level of clutter density was reached (Fig. 1B1–B3). Both the prey's and the predator's visual range was limited by the presence of occlusions. If an occlusion existed along a line between the prey and the predator then they could not see each other.

For this particular dynamic spatial navigation task, habit-based action selection exploited past action sequences that resulted in survival for a given environment. In real life, this would occur through trial and error over the lifetime of the animal or over evolutionary time; here following past practice (Daw et al., 2005) we obtain these trajectories from plan-based action selection. An action from this set was probabilistically chosen based on prior implementation outcomes. In other words, an action sequence was more likely to be chosen if it reliably resulted in survival when blindly followed

(Fernández & Veloso, 2006). In contrast, with plan-based action selection, within the imagination of the prey each virtual action was evaluated based on the virtual action's possible outcomes. The prey generated action sequences in imagination using the partially observable Monte-Carlo planning algorithm, which combines a sample-based approach to belief state update and to the tree of decisions the prey has at each state (Silver & Veness, 2010).

Results

Under pure plan-based action selection, in environments with very little clutter (entropy < 0.4), the predator speed and pursuit strategy restricts the prey's survival rate (Fig. 1A). As entropy increases to midrange levels (entropy 0.4–0.6), the prey's survival rate reaches its maximum (Fig. 1A). As entropy further increases (entropy > 0.6), both the effective size of the environment, and the number of possible escape routes to the safety position decreases. This in turn causes survival to be dependent on both the initialization of the environment, and the predator start location ($n = 5$) (Fig. 1A, B3).

Notably, if we look at all the trials in which the prey succeed in reaching the safety position to create the set of successful policies, referred to as "success paths", we observe that in both low and high entropy environments success paths are highly stereotypical (Fig. 1B1, B3). In low entropy environments these emergent successful action sequences resemble the wall-following behavior—or thigmotaxis—commonly observed in rodents in open-field tests. In high entropy environments, the amount of clutter constricts the profusion of success paths to one or two. The low spread of success paths in both low and high entropy environments enables habit-based action selection to perform at a level that is statistically indistinguishable from plan-based action selection (Fig. 1A). In contrast, at midrange levels of clutter the prey's survival strategy becomes less stereotyped, indicated by the increase in the number and spread of viable paths (Fig. 1B2). The spatial distribution of occlusions in these environments enables the prey that uses plan-based action selection to exhibit complex and flexible behaviors that strategically deploy occlusions to escape from the predator. This in turn causes habit to perform much worse than planning, since actions are not re-valuated under habit-based control (Fig. 1A).

The stereotypy seen in the resulting policy of the prey in low entropy environments suggests that the predator will employ a competing strategy that is similarly dependent on environmental connectivity. Complementary to our analysis of prey success paths, we quantified success paths for the predator by compiling a set of policies that resulted in successful prey capture (Fig. 1B4). Interestingly, these trajectories are similar to those observed in pursuit tasks in open environments with primates (Yoo, Piantadosi, & Hayden, 2018), and seem to arise as a result of easy access to predicted prey locations.

To formalize this pattern, we quantified the connectedness of the environment cells through eigenvector centrality (eigencentrality), which represents the weighted sum of direct connections through actions to and from a cell, as well as indirect connections of every length (Bonacich, 2007) (Fig. 1C). The stereotypical success paths employed by a prey that forward simulated 5000 states ahead are along cells of low eigencentrality. Conversely, the predator success paths, independent of its initial location, are more spatially distributed, and frequently occupy cells of high eigencentrality, which allows easy transitions to neighboring regions.

The increased spread of success paths in environments with mid-levels of clutter, and the corresponding emergence of complex behaviors, appears to be related to the distribution of eigencentrality. Unlike low entropy environments, which have a region of high eigencentrality in the center that tapers off in all directions away, midrange entropy environments exhibit adjacent clusters of highly and poorly connected regions (Fig. 1C, D). In such environments, the clustered nature of the eigencentrality forces the prey to transition from a relatively protected area to an open area. It is at this transition region that planning becomes imperative, since the prey has to account for the current predator location with respect to the occlusions to safely navigate exposed regions to the safety position (Fig. 1D). Preliminary support for this hypothesis is found in the pattern of nonlocal hippocampal spatial representations that sweep in front of rodents at high-cost choice points (Johnson & Redish, 2007), where there is a sharp change in eigencentrality (Fig. 1E). Given the success of pure habit-based action selection in both low and high entropy environments, we implemented a hybrid controller that uses habit-based action selection in low eigencentrality regions, and switches over to plan-based action selection at transition points to high eigencentrality regions.

At the start of an episode the control was initially given to habit. After each habit-based action (execution from a chosen policy (see Methods)), the prey compared the eigencentrality and the gradient of the eigencentrality at its current location to the next location that the policy prescribes. If both the eigencentrality and the eigencentrality gradient increased along the habitized action sequence, the control was transferred over to planning where the prey forward simulated 5000 future states prior to choosing an action. During plan-based action selection, similar to habit-based control, the prey compared the eigencentrality and the gradient of the eigencentrality at its current location to all the possible subsequent locations (e.g., not an occlusion, and not a wall). Switching to habit-based control occurred if the eigencentrality decreased and the absolute value of the gradient of eigencentrality increased.

Using the hybrid controller, in environments with low spatial eigencentrality clustering (low and high entropy) control was rarely transferred over from habit to planning (Fig. 1F). Consistent with our previous findings (Fig. 1A), in these environments, pure habit-based, pure plan-based, and hybrid control methods performed similarly (Fig. 1G; One way ANOVA $P = 0.71$). In environments that had high spatial clustering of eigencentralities (mid entropy), plan-based action selection was engaged more often (Fig. 1F). In these environments, pure habit-based action selection performed much worse than

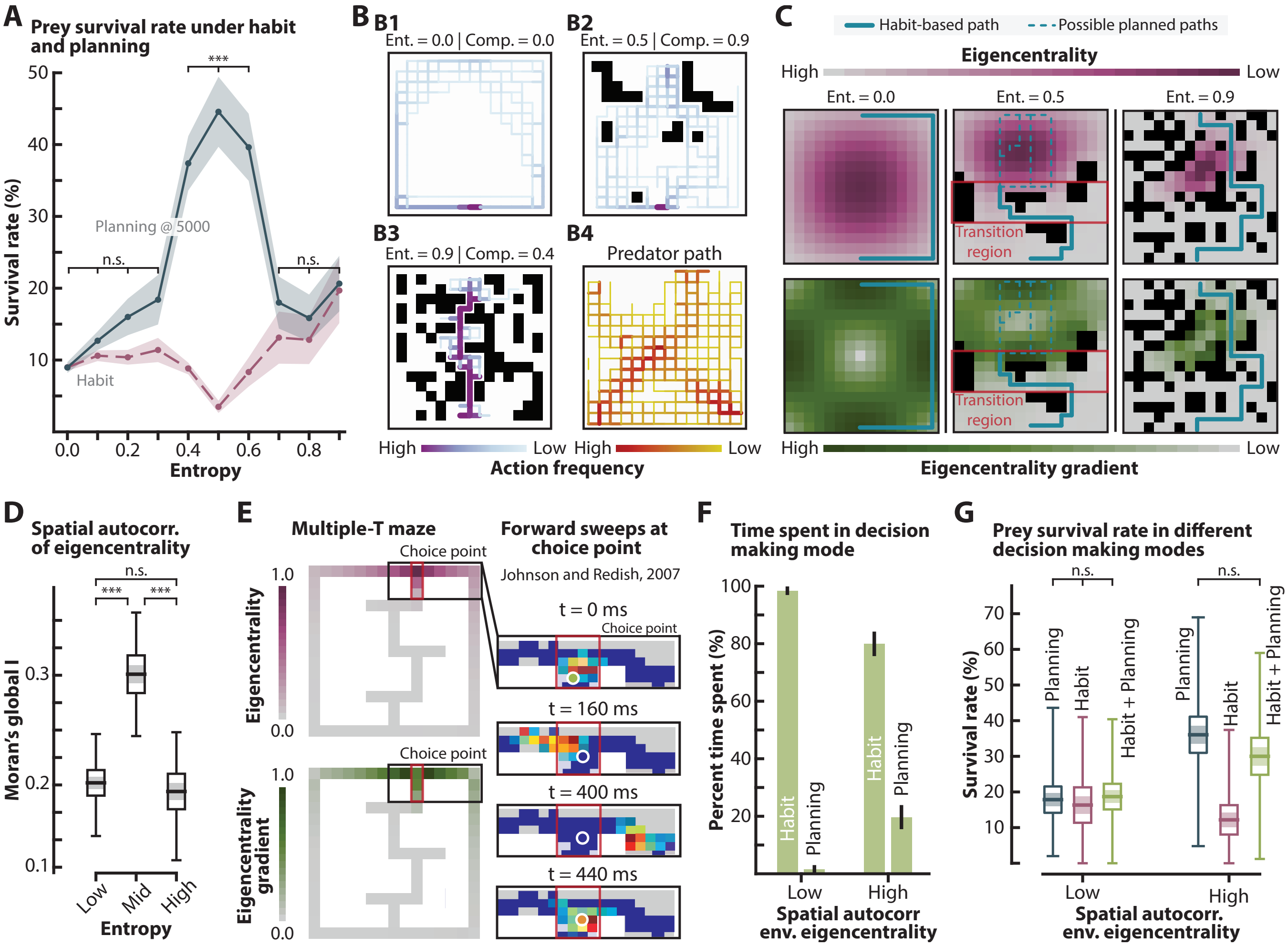

Figure 1: **(A)** Mean $\pm$ s.e.m ($n = 20$) of survival rate as a function of clutter level at 5000 states forward simulated (teal solid line). Mean $\pm$ s.e.m ($n = 20$) of survival rate as a function of clutter level for prey that rely on habit-based action selection (pink dashed line). There is no significant difference in survival rate between prey that uses habit and plan-based action selection at low (0.0-0.3) and high (0.7-0.9) entropy (One-way ANOVA Low: $P = 0.15$, High: $P = 0.52$, Mid: $P < 10^{-8}$). **(B1–B3)** Heatmaps of all action sequences taken by the prey that resulted in prey survival at 5000 states forward simulated, with color density proportional to action frequency. **(B4)** Heatmap of all action sequences taken by the predator that resulted in predator success (capture of prey), with color density proportional to action frequency. **(C)** Example environments and their eigencentralities and eigencentrality gradients. Color density of each metric proportional to the metric. Transition region from low to high eigencentrality based on change in gradient and change in eigencentrality value are shown by the red box. In our hybrid 'planning+habit' strategy, this transition region corresponds to a change in behavioral strategy from habit based (solid teal line) to planning (dashed line). **(D)** Spatial autocorrelation (global Moran's I) of the environment eigencentrality. Higher spatial autocorrelation indicates that the low and high values of environment eigencentrality are more spatially clustered (Mann Whitney U test with Bonferroni correction Low-Mid: $P < 10^{-6}$, Mid-High: $P < 10^{-5}$, Low-High: $P > 0.05$). The horizontal line corresponds to the mean, the shaded regions correspond to the s.e.m., and the boxes correspond to the 95% confidence interval of the mean. The line extending from the box depicts the range of the data. **(E)** Multiple T-maze overlaid with eigencentrality and eigencentrality gradient. Color density proportional to the metric. Red box indicates the "choice" point where the rat pauses. Johnson & Redish, 2007 showed that the neural representation (reconstruction on the right) moved ahead of the animal (white circle) while it paused at the choice point. **(F)** Average percent time spent in decision making regime (habit vs planning) when environments are grouped based on their spatial autocorrelation of eigencentrality. Low corresponds to the bottom 25%, which is largely made up low and high entropy environments. High corresponds to the highest 75%, which is largley made up mid-entropy environments. The error bars indicate $\pm$ s.e.m. of percent time ($n_{\text{low}} = 54$, $n_{\text{high}} = 50$). **(G)** Survival rate for a prey that uses planning (blue), uses habit-based action selection (pink), and uses hybrid control (green) based on environment eigencentrality. Environment grouping same as **F**. Plot representation as in **D**.

pure plan-based action selection (Fig. 1A, G). However, the hybrid strategy that engaged planning when transitioning from a low eigencentrality region to a high eigencentrality region significantly outperformed pure habit-based action selection, and showed no significant difference in performance when compared with pure plan-based action selection (Fig. 1G; Mann Whitney U test with Bonferroni $P > 0.05$). Notably, using a hybrid control strategy resulted in 75%-85% of the time being spent in habit-based action selection, which led to only a 9% decrease in survival rate when compared to the survival rate obtained from pure plan-based action selection.

Conclusion

Prior work on prospective coding in the hippocampus indicates that the forward sweeping of spatial representations occurs at important decision points when the reward contingencies are uncertain (Johnson & Redish, 2007). The predator-prey model we have used for this study is just a subset of this broader phenomena where the animals have uncertainty about where the reward is located. Within this model, the predator takes on the role of an unpredictable, sometimes unobservable aversive stimulus that has to be avoided. Our results indicate that in dynamic environments, after consolidating long successful action-sequences, selecting actions by using a habit-based system would enable animals to succeed in getting to the safety location in environments with either low or high levels of clutter. In environments with midrange levels of clutter, while pure habit-based control fails, we show that planning throughout the entire episode is not necessary. A controller that arbitrates between habit- and plan-based action selection that exploits the connectivity of the environment performs just as well as pure planning.

Prior research has hypothesized that control is transferred over from planning to habit based on uncertainty in state values (Daw et al., 2005). Our findings fit within this literature, since the prey's uncertainty is proportional to the regional openness and reachability—as quantified by eigencentrality. While our work has primarily focused on switching from plan- to habit-based action selection (and visa versa) in spatial domains, eigencentrality as a concept has broader applicability. Planning, similarly is not purely spatial. Given that our framework relies on formalizations that extend beyond our current application, it would be interesting to examine whether transitions from habit- to plan-based action selection based on changes in eigencentrality extend to non-spatial contexts.

Acknowledgments

This work was funded by NSF Brain Initiative EECS-1835389.

References

Adhikari, A., Topiwala, M. A., & Gordon, J. A. (2010). Synchronized Activity between the Ventral Hippocampus and the Medial Prefrontal Cortex during Anxiety. *Neuron*, *65*(2), 257-269.

Benchenane, K., Peyrache, A., Khamassi, M., Tierney, P. L., Gioanni, Y., Battaglia, F. P., & Wiener, S. I. (2010). Coherent Theta Oscillations and Reorganization of Spike Timing in the Hippocampal-Prefrontal Network upon Learning. *Neuron*, *66*(6), 921-936.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*, *29*(4), 555–564.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Elliott, J. P., Cowan, I. M., & Holling, C. (1977). Prey capture by the african lion. *Canadian Journal of Zoology*, *55*(11), 1811–1828.

Fernández, F., & Veloso, M. (2006). Probabilistic policy reuse in a reinforcement learning agent. In *Proceedings of the fifth international joint conference on autonomous agents and multiagent systems* (pp. 720–727).

Hedenström, A., & Rosén, M. (2001). Predator versus prey: on aerial hunting and escape strategies in birds. *Behavioral Ecology*, *12*(2), 150–156.

Ito, H. T., Zhang, S.-J., Witter, M. P., Moser, E. I., & Moser, M.-B. (2015). A prefrontal–thalamo–hippocampal circuit for goal-directed spatial navigation. *Nature*, *522*(7554), 50.

Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, *27*(45), 12176–12189.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*(5), e1002055.

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 1609-1617.

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*.

Schmidt, B., Duin, A. A., & Redish, A. D. (2019). Disrupting the medial prefrontal cortex alters hippocampal sequences during deliberative decision-making. *Journal of Neurophysiology*.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Silver, D., & Veness, J. (2010). Monte-carlo planning in large POMDPs. In *Advances in neural information processing systems* (pp. 2164–2172).

Sutton, R., & Barto, A. G. (2018). *Reinforcement learning: An introduction (2nd edition)* [Book]. Cambridge, Massachusetts: The MIT Press.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464.

Yoo, S. B. M., Piantadosi, S. T., & Hayden, B. Y. (2018). Monkeys predict trajectories of virtual prey using basic variables from newtonian physics. *bioRxiv*, 272260.