Connecting Exploration, Generalization, and Planning in Correlated Trees

Tobias Ludwig[1] **(tobias.ludwig@tuebingen.mpg.de),Charley M. Wu**[1,2]**, & Eric Schulz**[1]
[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2]Human and Machine Cognition Lab, University of Tübingen, Tübingen, Germany

Abstract

Human reinforcement learning (RL) is characterized by different challenges. Exploration has been studied extensively in multi-armed bandits, while planning has been investigated in multi-step decision tasks. More recent work added structure to bandits to study generalization. However, most studies focus on a single aspect of learning, making it hard to compare and integrate results. Here, we propose a generative model for constructing Correlated Trees, which provide a unified and scalable method for studying exploration, planning, and generalization in a single task. In an online experiment, we found that, when provided, people use structure to generalize and perform uncertainty-directed exploration, with structure helping more in larger environments. In environments without structure, exploration becomes more random and more planning is needed. All behavioral effects are captured in a single model with recoverable parameters. In conclusion, our results connect past research on human RL in one framework using Correlated Trees.

Keywords: multi-step decisions; correlated environments; exploration; generalization; planning; trees.

Introduction

Any agent placed in a sufficiently complex environment faces a similar set of challenges. If the goal is to maximize rewards, then the agent needs to balance between exploring unfamiliar options to gain information and exploiting options that are known to be good (Schulz & Gershman, 2019). Moreover, if the structure of rewards are predictable, then intelligent agents should use this structure to generalize from past knowledge to unseen options (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018), thereby speeding up learning by directing exploration towards promising options. Finally, in sequential decision problems, maximizing rewards requires planning the best sequence of actions rather than only selecting actions myopically (Daw, Gershman, Seymour, Dayan, & Dolan, 2011).

Given the importance of these challenges, it comes as no surprise that past research on human reinforcement learning (RL) has focused substantially on how people *explore* (Wilson, Geana, White, Ludvig, & Cohen, 2014; Speekenbrink & Konstantinidis, 2015), *generalize* (Franklin & Frank, 2020; Wu, Schulz, Garvert, Meder, & Schuck, 2020), and *plan* (Huys et al., 2015; Keramati, Smittenaar, Dolan, & Dayan, 2016). Yet past studies have commonly studied and designed models for a singular dimension of learning. However, in order to generalize our understanding of human RL more broadly, we would like to integrate our insights over whole classes of problems that assess how people explore, generalize, and plan.

In the current work, propose a method to bridge diverse paradigms and models of human RL. Specifically, we programmatically generate tasks with different depths (d), branching factors or breadths (b), and reward correlations (c) (Fig. 1A). Depth is mapped to planning because it determines the number of steps on any path. The branching factor modulates exploration by defining how many options need to be considered in any state. Note that we subsume the class of multi-arm bandits and multi-step planning tasks just by these two parameters. Finally, we add correlation between rewards to provide traction for generalization.

Our behavioral results show that people use structure to generalize about correlated rewards, explore in a directed fashion, and plan ahead by taking the rewards of next steps into account. These effects were all stronger in correlated reward environments. Participants' behavior was captured and reproduced well by a Bayesian model of generalization, exploration, and planning, whose parameter estimates were highly recoverable. Our model results revealed a trade-off between generalization and planning, as well as between directed and random exploration, depending on the reward structure. Taken together, our results connect past research on human RL in one coherent experimental and modelling framework, and pave the way for future investigations of generalization, exploration, and planning in complex environments.

Experiment: Correlated Trees

We use decision trees as a framework to study decision processes with an arbitrary number of steps. Each node is a state in the task associated with some reward. Participants were asked to make repeated trips from the root node to one of the terminal nodes, accumulating as much reward as possible over a fixed number of trials.

Generative model

The depth $d \in \{2,3,4\}$ of a tree determines the number of decisions the agent needs to make to reach a leaf node, and the branching factor $b \in \{2,3,4\}$ defines the number of possible options in each state. Each node generates noisy rewards, where the reward structure is determined by a correlation parameter $c \in \{0,1\}$, corresponding to random or structured rewards, respectively.

In the random reward condition ($c = 0$), nodes were sampled independently from a normal distribution $\sim \mathcal{N}(50,25)$. In the structured reward condition ($c = 1$), we defined a correlated reward structure such that nodes connected by an edge produced similar rewards. In addition to traversable edges of the tree (straight edges in Fig. 1B), we added non-traversable lateral edges between neighboring sibling nodes on the same level (rounded edges in Fig. 1B) to add correlations between options within each decision. Specifically, expected rewards for each node were sampled from a Gaussian Process ($\mathcal{GP}$) prior, parameterized by a diffusion kernel (Wu, Schulz, & Gershman, 2021; Kondor & Lafferty, 2002):

$$\mathbf{r} \sim \mathcal{GP}(\mu, \mathbf{K}), \qquad \mathbf{K} = \sigma^2 \cdot \exp(-c \cdot \mathbf{L}) \tag{1}$$

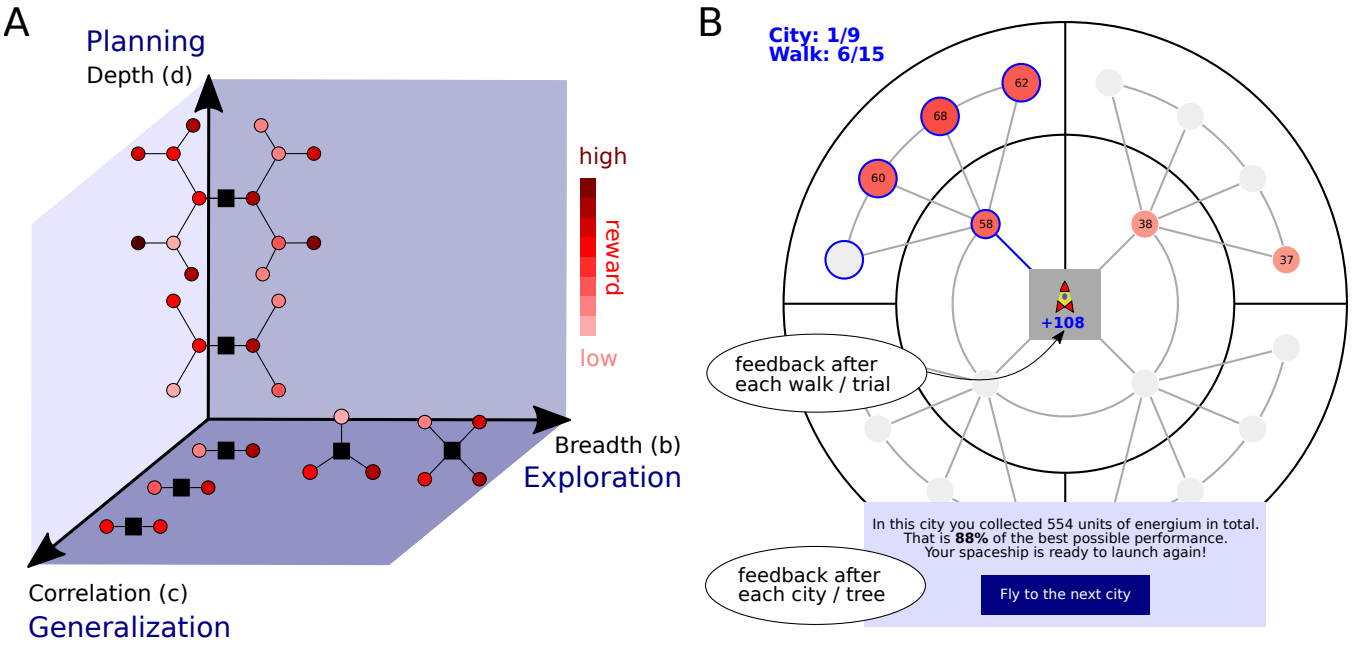

Figure 1: **A. Tree space**. Our space of Correlated Trees is defined by depth (d), branching factor (b), and correlation strength (c). **B. Experiment**. Example tree ("alien city") used in the experiment ($d = 3, b = 4, c = 1$). Nodes had to be uncovered trial by trial, by walking from the center square to a terminal node on the outermost level via the grey "streets". The lateral streets (arcs) introduce correlations between sibling nodes but were not traversable. In the example, the agent is currently in the 58-node and can transition to one of the four blue-bordered nodes, three of which have previously been visited (reddish) and one is unvisited (grey). Black lines serve to visually separate sub-trees.

The kernel $\mathbf{K}$ defines a covariance structure based on the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the degree matrix and $\mathbf{A}$ is the adjacency matrix of the graph. Again, we used a mean of $\mu = 50$ and a standard deviation of $\sigma = 25$. Since reward variance over paths increases with correlation strength, we sampled the uncorrelated trees such that they had the exact same path sums as a corresponding correlated tree, in order to balance attainable rewards in both conditions.

Methods

Participants and design. We ran an online study on Prolific ($N = 107$), where we manipulated reward structure between-subjects ($c = 0$ vs. $c = 1$) and manipulated tree structure within-subject ($b \in \{2,3,4\} \times d \in \{2,3,4\}$). Payment was performance-dependent and averaged 9.93 GBP per hour, with a median task duration of 27.66 min. Upon data inspection, we excluded 8 subjects who performed worse than 2 standard deviations below the mean score, plus one subject with data loss. Our final sample included 98 participants ($N_{c=0} = 48$ and $N_{c=1} = 50$; 41 female; mean age = 26.47).

Materials and procedure. Participants were given a cover story describing the trees as alien cities (Fig. 1B). Their task was described as collecting energy units (i.e., rewards) by repeatedly traveling from the center square to the outskirts (i.e., a terminal node in the outer level). The instructed goal was to maximize the total reward in each city, which could be achieved by finding and exploiting the best path.

Participants assigned to the random reward condition were explicitly told that no structure can be used to direct their search and all rewards were independent. Participants assigned to the correlated condition were instructed that there were meaningful similarities between nodes connected by a street (including the lateral connections). Reward observations included Gaussian noise $\sim \mathcal{N}(0,2)$. For each new city, all nodes were initially shown in grey, but upon clicking, they displayed a numerical label and color indicating the observed reward. All visited nodes remained visible, displaying the most recent observation. Participants were explicitly informed that rewards would not diminish as a consequence of sampling the same node twice (as energy units would be recharged between walks), which was reinforced during the tutorial and confirmed during a comprehension check. We also ensured that participants could never know if they found the best node already, because the maximum reward varied from city to city.

We denote each walk from the center to a terminal node as a *trial*. After each trial, participants were teleported back to the center, where they were shown the cumulative reward for the trial (for 2 sec), and then started anew. There were 15 trials per city and 9 cities in total (all combinations of b and d). The number of unique paths varied between 2 and $4^3 = 64$. We pre-generated 10 tree versions of each size, which were randomly assigned to participants in the task. Each participant completed all the tree sizes once (with no repetitions). After each city, participants were shown their average reward as a percentage relative to the best possible reward, and they received bonus payments relative to their overall score.

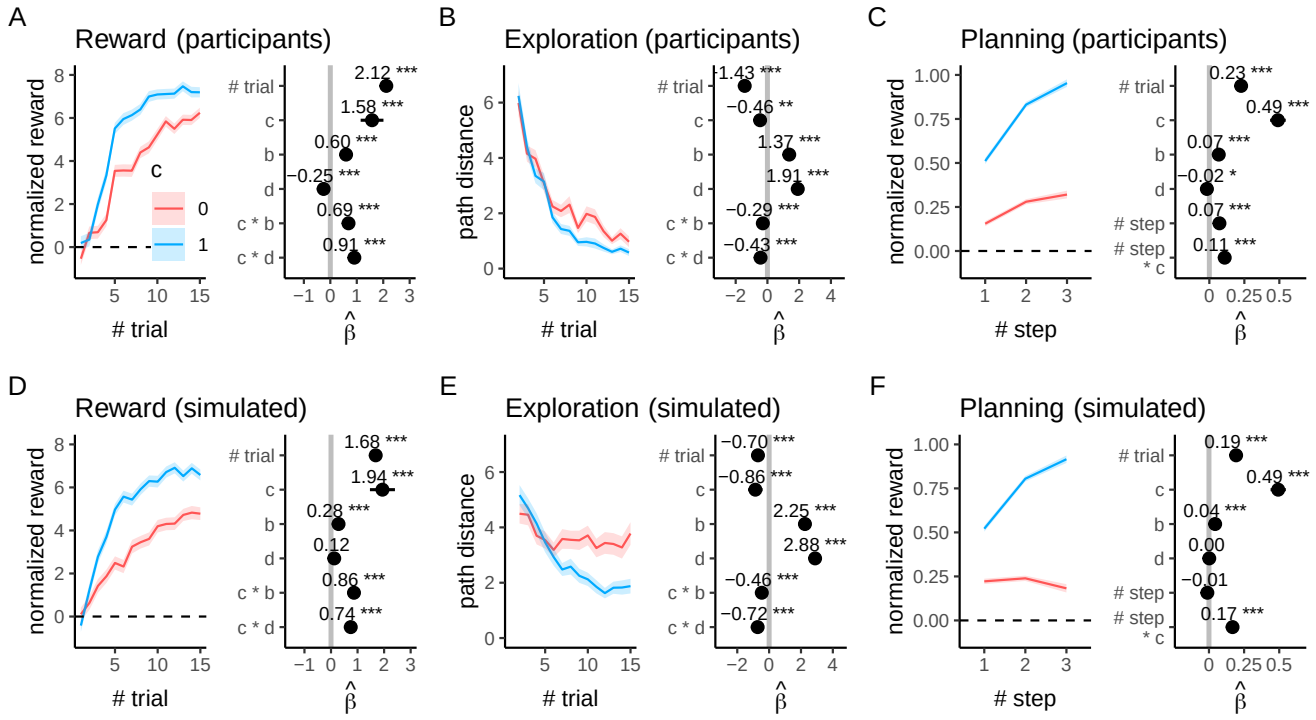

Figure 2: **behavioral results. A** Path reward of each trial corrected for chance (zero). **B** Path distance between consecutive trials. **C** Amount of reward for a single step within a trial (corrected for statistics within a tree-level; chance is zero). Next to each panel, we show corresponding regression weights $\hat{\beta}$ from a mixed linear model (* $p < .05$, ** $p < .01$, *** $p < .001$). The bottom row (panels **D**, **E**, **F**) mirrors the upper one, but shows data simulated by our full model (see Modelling results).

Results

Behavioral results

A main comparison of interest was the between-subject manipulation of reward correlations ($c = 0$ vs. $c = 1$). A secondary aspect was the scaling of behavior across different tree sizes (within-subject). Thus, Figure 2A-C focuses on plotting the effect of reward correlations on performance, exploration, and planning (lines), while reporting the effects of tree structure and their possible interactions as regression weights (dot plots). We used linear mixed effect regressions with participant-specific random intercepts, where all predictors were z-standardized.

Performance. Figure 2A shows that performance in both conditions improved over time, but less quickly and with a lower asymptote for $c = 0$. The linear model confirmed a significant main effect of c ($\hat{\beta} = 1.58 \pm 0.22, p < .001$), and revealed strong interaction effects of c with both branching factor b ($\hat{\beta} = 0.69 \pm 0.10, p < .001$) and depth d ($\hat{\beta} = 0.91 \pm 0.10, p < .001$). This suggests subjects made use of the reward structure, and benefited even more from it in larger environments. Note that the random reward environment was not more difficult *a priori*, since we controlled for chance and the conditions were matched to equal attainable rewards.

Exploration. Next, we looked at how participants explored, hypothesizing they would begin by exploring diverse paths to get a broad overview, before narrowing down their search to exploit the most promising paths. This form of strategic search should work better in the correlated condition, given the increased traction that generalization provides for exploration. We constructed a behavioral measure of exploration called the "path distance" between consecutive trials, defined based on the depth at which paths forked and how far apart they forked.

Figure 2B shows how path distance decreased over trials, but remained larger for the $c = 0$ condition ($\hat{\beta} = -0.46 \pm 0.17, p = .008$). Path distance also increased with the size of the tree — trivially — since this would even happen under a random policy. Nonetheless, the negative interaction of c with b ($\hat{\beta} = -0.29 \pm 0.09, p < .001$) and c with d ($\hat{\beta} = -0.43 \pm 0.09, p < .001$) suggested that larger tree sizes did not increase exploration much in correlated cities. One interpretation is that as a result of successful generalization, large parts of the tree could be avoided due to expectations of poor rewards, allowing for less but more efficient exploration.

Planning. Lastly, we were interested in how far people looked ahead when planning their walks. We hypothesized that more look-ahead would yield higher rewards on later steps of the walk. In contrast, following rewards myopically would result in higher rewards in earlier steps. Figure 2C

shows reward as a function of steps, averaged over all walks (corrected for chance using z-scoring w.r.t. the statistics of the rewards within a tree-level). We excluded 1-step ($d = 2$) trees to focus only on genuine multi-step decisions. Our results suggest people looked ahead in both conditions, with a significant main effect of c ($\hat{\beta} = 0.49 \pm 0.03, p < .001$), as well as an interaction of c with the step number ($\hat{\beta} = 1.10 \pm 0.01, p < .001$). This implies that exploiting structure helps, especially for finding higher distant rewards.

Modelling results

We model participant behavior using a single model combining generalization, exploration, and planning components. A Gaussian process (GP) regression model with a diffusion kernel provides a method of generalization (Wu et al., 2021), Upper Confidence Bound (UCB) sampling provides a mechanism for performing uncertainty-directed exploration (Wilson et al., 2014; Schulz & Gershman, 2019), and temporal discounting of more distant nodes provides a means to describe myopia in planning.

We first describe how a GP can be used to generalize rewards from past observations onto unobserved nodes. We use the same diffusion kernel $\mathbf{K}$ from Eq. 1, but use α in place of c as a free parameter, since the true generating c is unknown to the subjects. With every new observed reward y at a node x, the GP model updates its mean m and variance v predictions for each node according to the following posterior:

$$m'(x) = m(x) + \mathbf{K}_{xX}(\mathbf{K}_{XX} + \sigma_\varepsilon^2 \mathbf{I})^{-1}(y - m(X)) \quad (2)$$

$$v'(x) = v(x) + \mathrm{diag}\left(\mathbf{K}_{xX}(\mathbf{K}_{XX} + \sigma_\varepsilon^2 \mathbf{I})^{-1}\mathbf{K}_{Xx}\right) \quad (3)$$

where $\mathbf{K}_{xX}$ denotes evaluation of the kernel at node x paired with all other nodes X. The GP was initialized with $m_0 = 50$, corresponding to the true mean reward μ in the environment, and $\sqrt{v_0} = 30$ was slightly higher than the true sd ($\sigma = 25$). The noise variance was the same as in the task $\sigma_\varepsilon^2 = 2$.

Next, we define node utility u using UCB as a weighted sum of the estimated mean and variance of each node:

$$u(x) = m(x) + \beta\sqrt{v(x)}. \quad (4)$$

The non-negative β parameter controls the amount of uncertainty bonus for variance-directed exploration, such that $\beta = 0$ would correspond to a mean-greedy policy.

Since the task is not to find the best node but the best path, we define the utility of a path as a sum over discounted node utilities, where the discount parameter γ controls how much the agent values future prospects:

$$U(x_1, x_2, \dots) = \gamma^0 u(x_1) + \gamma^1 u(x_2) + \dots \quad (5)$$

$\gamma = 0$ corresponds to myopic behavior, which only takes into account immediate rewards, while $\gamma = 1$ would equally value each node along the path.

Lastly, we apply a softmax function to transform these path utilities into corresponding choice probabilities

$$p_i = \frac{\exp(U_i/\tau)}{\sum_j \exp(U_j/\tau)}, \quad (6)$$

where higher values of the temperature parameter τ corresponds to more random exploration.

In summary, our model describes generalization via the diffusion parameter α, uncertainty-directed exploration via the UCB weight β, and accounts for myopia via the discount factor γ. The softmax temperature parameter τ introduces a second, more random form of exploration.

Model fitting

We fit the model to each subject individually using a maximum likelihood approach, yielding the most likely parameter set $(\alpha, \beta, \gamma, \tau)$ and the model likelihood $\mathcal{L}$ under these parameters. Additionally, we fit various lesioned versions of our model by systematically removing components via fixed parameters. We compare models using $BIC = -2\log\mathcal{L} + k \cdot \log(n)$ where k is the number of parameters (4 for the full model, 0 for random policy) and $n = 15 \cdot 9 = 135$ is the number of trials. For intuition, we report goodness of fit using a pseudo-R^2 measure:

$$R^2 = 1 - \frac{BIC(model)}{BIC(random)} \quad (7)$$

as an interpretable comparison to a random baseline. Intuitively, $R^2 = 0$ indicates chance level predictions and $R^2 = 1$ is a perfect model.

Model comparison. How well does our model capture participant data? Taking a random policy as a baseline (corresponding to $R^2 = 0$), our model fits the correlated condition much better than the uncorrelated condition (mean $R^2 = .394$ vs. $R^2 = .255$; 2-sample t-test, $t(95) = -5.021, p < .001$; Fig. 3A). This makes sense, because behavior in the uncorrelated condition was expected to be more random and less predictable.

We then performed several comparisons to lesioned models, defined by fixing one of the model parameters, to test the importance of the generalization, directed exploration, and planning components. Figure 3B shows the corresponding difference in R^2 to the full model, where negative values indicate superiority of the full model.

We first lesioned the ability to generalize by fixing $\alpha = 0$. We expected this to specifically impact the correlated reward condition, but be quite adaptive for the uncorrelated condition (where $c = 0$). Unsurprisingly, the $\alpha = 0$ model performed worse in the correlated condition (paired t-test, $t(49) = -5.20, p < .001$), but better in the uncorrelated condition ($t(47) = 5.82, p < .001$). Next, we lesioned directed exploration by setting $\beta = 0$, which produced worse predictions in both conditions ($c = 0 : t(47) = -2.91, p = .006$; $c = 1 : t(49) = -6.3, p < .001$), but with a much larger effect in the correlated condition. Lastly, we use $\gamma = 1$ as a lesioned form of temporal discounting, in which all nodes are treated equally. We find that there was no difference in comparison to the full model for the correlated condition ($t(49) = -0.28, p > .05$), which is consistent with γ estimates

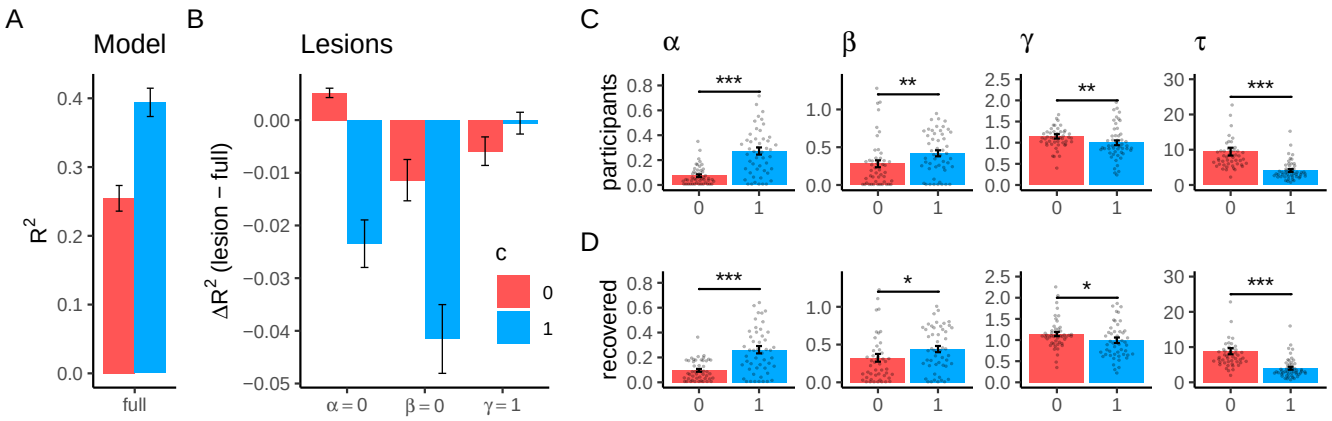

Figure 3: **Modelling results. A**. Goodness of fit reported as a pseudo-R^2. **B**. Lesion analysis, where we compared model variants with a fixed parameter against the full model in Panel A. Negative values indicate worse fits than the full model ($\Delta R^2 = R^2_{lesion} - R^2_{full}$). **C**. Parameter estimates of the full model on subject data. **D**. Recovered parameters based on fitting our models to simulated data (using parameter estimates from panel C). Bars show mean estimates by condition and errorbars show standard errors. Superimposed dots show the values of single subjects and asterisks indicate significance between groups (Wilcoxon signed-rank test, * $p < .05$, ** $p < .01$, *** $p < .001$).

in the full model (Fig. 3B) being quite close to 1. However, the $\gamma = 1$ lesion performed worse for uncorrelated rewards ($t(47) = -2.18, p = .034$), suggesting that the best fit for γ in this condition is different from 1 — in fact it is larger, as described below.

Simulated behavior. As an additional check on the reliability of our full model, we simulated behavior using participant parameter estimates (Fig. 2D-F). By and large, all effects were reproduced qualitatively. Quantitatively, asymptotic performance is a bit lower, reflecting that our model is not perfect. Notably, there is profoundly less decay in exploration distance for the $c = 0$ condition. This suggests that humans might use a dynamic exploration schedule that is not captured by our model (e.g., changing τ or β over trials). Furthermore, the increase in reward per step (Fig. 2F) is less steep for the $c = 0$ condition than in the subject data.

Parameter estimates. We now focus on interpreting the parameters of our model, which we compare across reward conditions. Figure 3C shows the parameter fits on participant data, while the same effects display in the recovered parameters (Fig. 3D). Firstly, we found that participants used structure when it was provided. In line with our lesion analysis, participants generalized more in the correlated than in the uncorrelated condition (Wilcoxon signed-rank test, $W = 438, p < .001$). Notably, the estimated α in the correlated condition was still significantly below the true $c = 1$, which echos past evidence of under-generalization in spatially- (Wu et al., 2018) and graph-correlated bandits (Wu et al., 2021).

Secondly, β estimates were slightly higher ($W = 818, p < .007$) in the correlated condition. This may be because through generalization, directed exploration is more efficient. In contrast, in the uncorrelated condition, variance estimates are not meaningfully informed by neighboring nodes, and thus a lower uncertainty bonus makes sense. For a similar argument, we see the opposite trend in τ, driving random exploration, which was higher in the uncorrelated condition ($W = 2087, p < .001$).

Lastly, we found higher γ estimates in the uncorrelated condition ($W = 1567, p < .009$), whereas the lesion analysis revealed that a fixed $\gamma = 1$ predicts better in the correlated condition. Curiously, participants in the uncorrelated condition were better fit by a γ slightly above 1 (one sample t-test, $t(47) = 2.87, p = .006$). Since we did not constrain the model to using $0 \leq \gamma \leq 1$, as is common in RL models, here $\gamma > 1$ suggest that more weight was placed on more distant rewards. This intuition aligns with the natural statistics of the task: the most rewarding nodes are more likely positioned in the last step, since the outermost level contains the most nodes. While we corrected for this in the behavioral analysis (Fig. 2C/F), the same was not possible for the models. In sum, participants were oriented towards high rewards at the outer level in the uncorrelated condition, but valued all nodes equally in the correlated condition. We interpret this as a trade-off between generalization and planning: if search is not guided by correlation structure, more look-ahead is needed.

Discussion

We have proposed correlated trees as a scalable environment to study multiple aspects of human decision making in a single task. Specifically, we combined generalization, exploration, and planning in a single model, and fit it to behavior, where we found more generalization and directed exploration when correlation structure was available. The benefit from correlation was especially high in larger environments. On the other hand, if no correlation was provided, exploration became more random and more planning was needed. Thus, we observed a different trading-off, depending on the structure and size of the environment.

Limitations and future directions. Our paradigm allows us to study exploration in three dimensions, across depth, breadth, and reward correlations. The interaction of depth and breadth introduces an additional trade-off on top of the classical exploration-exploitation dilemma (Moreno-Bote, Ramírez-Ruiz, Drugowitsch, & Hayden, 2020). For instance, we see a similar decay of path distance as in Figure 2B, even if we exclude exploitative (i.e., mean-greedy) trials. This suggests that subjects start by sampling very broadly (in breadth), and then narrow their search towards a promising direction in depth. However, in the current version of the experiment the two dimensions are not perfectly separable, as the subjects were forced to descend the whole path in a given trial, and walking backwards or side-ways was prohibited.

Studying the same task in environments of different sizes should also tell us a lot about how human exploration, generalization, and planning scales with complexity. For example, we saw behaviorally that generalization supports exploration in large environments because large parts of the tree can be avoided. It would be interesting to uncover these scaling effects also on a modelling level. However, here we were limited to fitting a single model for each participant, which aggregated across differently sized trees. Additionally, our present task had the limitations that a) we could only scale to a maximal breadth and depth of 4 (corresponding to 64 paths) for visual reasons, and b) the number of trials was kept constant at 15 regardless of the size of the environment. The latter point bounded the amount of exploration that was possible in large trees, whereas small trees could be exhaustively explored. Since Wilson et al. (2014) have shown that the number of trials ("horizon") has a crucial influence on how people explore, future tasks should take this into account when comparing how behavior scales across environment sizes.

Our current model uses root planning to calculate the path utilities for each path starting at the root of the tree. Yet, this is rather unlikely for human planning, because it requires exhaustive computations for every possible path (here, up to 64). In contrast, people are likely able to reduce the search space using heuristics (e.g., by visual search for good nodes). An alternative to root planning would be an online strategy, in which the agent sequentially plans each node along the way. Whereas a myopic root planner would only maximize utility on the first node (picking later ones at random), online planning would allow for queuing multiple myopic decisions. However, cursory analysis of reaction times suggested most of decision time was spent on the root note in both conditions.

Another way in which we studied a rather limited sense of planning, is that any thinking-ahead was a mere looking-ahead. In the current setup, subjects always saw the whole space of possible paths, without the need to retrieve past experiences from memory. Moreover, we did not account for stochastic transitions, which are an important feature of tasks used in the planning literature (Daw et al., 2011). Future work could introduce such transitions in the task and obscure visited nodes so to require a more model-based value learning.

Lastly, most natural environments are not necessarily tree-shaped. Real cities, for instance, have no walls between neighborhoods and allow for walking laterally, with more than one path leading to any place. Also, correlation structures are more nuanced than the simple diffusion-based correlation employed here. In particular, there might be negative correlations, which we ignored so far.

Nevertheless, trees provide an interesting intermediate environment bridging the gap between studies in bandits and state-full Markov Decision processes (MDPs) in terms of complexity and realism (Brändle, Binz, & Schulz, 2021). Future work could extend the idea of a unifying framework for exploration, generalization, and planning to general MDPs.

Conclusion

We made a case for studying exploration, generalization, and planning jointly, proposing correlated trees as a scalable environment for doing so. We tested the task in a behavioral experiment, and fitted a model that revealed trade-offs between generalization and planning as well as between random and directed exploration.

References

Brändle, F., Binz, M., & Schulz, E. (2021). Exploration beyond bandits.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Franklin, N. T., & Frank, M. J. (2020). Generalizing to generalize: humans flexibly switch between compositional and conjunctive structures during reinforcement learning. *PLoS computational biology*, *16*(4), e1007720.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., ... Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098–3103.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences*, *113*(45), 12868–12873.

Kondor, R. I., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th international conference on machine learning* (Vol. 2002, pp. 315–322).

Moreno-Bote, R., Ramírez-Ruiz, J., Drugowitsch, J., & Hayden, B. Y. (2020, August). Heuristics and optimal solutions to the breadth–depth dilemma. *Proceedings of the National Academy of Sciences*, *117*(33), 19799–19808.

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

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, *7*(2), 351–367.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2020). Similarities and differences in spatial and non-spatial cognitive maps. *PLoS computational biology*, *16*(9), e1008149.

Wu, C. M., Schulz, E., & Gershman, S. J. (2021). Inference and search on graph-structured spaces. *Computational Brain & Behavior*, *4*, 125–147.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature human behaviour*, *2*(12), 915–924.