Exploring human learning and planning in grid navigation with arbitrary mappings.

Carlos Velázquez (cavargas@princeton.edu)
Department of Psychology, Princeton University,
Princeton, NJ 08540 USA

Jordan A. Taylor (jordanat@princeton.edu)
Department of Psychology, Princeton University,
Princeton, NJ 08540 USA

Abstract

From learning to play video games to using novel tools, humans are able to acquire a variety of complex mappings between their actions and arbitrary outcomes. In addition, once they have learned such mappings, they often have to use them sequentially to achieve goals, i.e., planning. In this work, we study how the learning of a novel mapping interacts with planning in the context of grid navigation. In order to do so, we developed a computer-based game where subjects have to move a cursor from start to target locations using the keys of their keyboard. Importantly, to more closely resemble the complexity of the mappings that people acquire in their lives, the cursor movement was determined by a non-trivial rule inspired by the movement of the Knight chess piece. In Experiment 1, we show that participants were able to improve their performance in our task, though not always arriving to the targets optimally. Additionally, we explored different classes of cognitive models and found that a model that includes Bayesian mapping-learning, path search and habit formation components best described the data. In Experiment 2, we asked whether breaking down the task into its mapping-learning and planning components could improve participants' performance. Indeed, we found that exposing participants to the mapping component of the task without having to plan, provides a performance improvement when exposed to the full task later. Crucially, this improvement does not occur if subjects are exposed to the planning component of the task prior to doing it fully. Overall, these results suggest that in order for planning processes to be effectively deployed, the mapping of actions should be learned first.

Keywords: planning; Bayesian learning; motor learning.

Introduction

During their lifetime, humans can develop a wide and complex repertoire of skills such as dancing, swimming, riding a bicycle, playing musical instruments or playing video games. The intricate nature of these activities has made their scientific study equally complex. On the one hand, humans have to figure out the motor commands that lead to the desired outcomes. For example, which configuration of the hand produces a given chord on the guitar or what button presses make a character jump or walk in a video game? The formation of this mapping is arguably one of the most important steps to develop a skill (Fitts & Posner, 1967; Adams, 1971; Ackerman, 1988; Newell, 1985, 1991). Surprisingly, how the brain learns these novel and, often arbitrary, mappings remains poorly understood. A second challenge arises as the majority of complex skills are extended in time, involving sequencing together a set of actions with the mapping to accomplish goals. For example, a given combination of chords is necessary to generate a song, and sequences of button presses make players navigate through different levels of a video game. The dependence of goals on the concatenation of actions gives rise to one of the key processes of human cognition: planning (Hunt et al., 2021). In the current work we aim to understand how the learning of a sensorimotor mapping interacts with planning to acquire complex skills. Additionally, we aim to provide a computational account of the potential mechanism of interaction.

Developments on motor learning and planning research currently have no unifying framework on this matter. On the one hand, the acquisition of a novel mapping has been studied in sequence learning tasks where people have to learn what action to take, normally key presses, when arbitrary visual stimuli are presented on a computer screen (Balsters & Ramnani, 2011), however, there is generally no overarching goal towards which people can freely use the mapping, i.e., choosing their own sequence of actions, such as in video games. Additionally, in most sequence learning tasks, the sequence to be learnt is specified by the experimenter (Korman, Raz, Flash, & Karni, 2003; Kami et al., 1995). This limits considerably the planning aspect of the tasks. One exception is the work on grid navigation (Fermin, Yoshida, Ito, Yoshimoto, & Doya, 2010; Bera, Shukla, & Bapi, 2021b, 2021a), where people have to move a cursor to target locations on a grid using keys with an unknown movement-mapping. In this task the learned mapping can be freely manipulated to achieve the goal of the task, potentially involving planning processes. However, the computational mechanism of how people succeed in scenarios like this is still unknown.

On the other hand, planning research has focused on how humans or artificial agents can maximize their future rewards following a series of actions. Reinforcement learning algorithms have been developed to achieve this goal, particularly model-based algorithms (Hunt et al., 2021). In these formalisms, a model of the environment is assumed to be known and used to simulate the outcomes of future decisions with the goal of choosing the ones that are the best. Great progress has been made on how to make planning a tractable computation (Schrittwieser et al., 2020; Hunt et al., 2021) and also on how humans actually plan (van Opheusden et al., 2021). However, in this research, the actions on which planning operates (the mapping) are often assumed to be known. This is at odds with real life learning scenarios where humans have

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

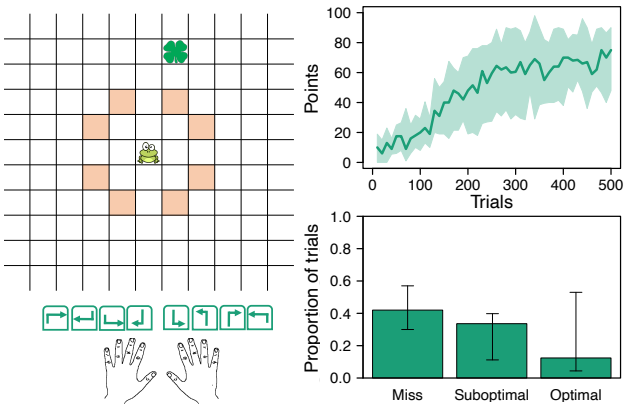

Figure 1: Experiment 1 task and behavioral results. Left: Participants moved a cursor (frog) from start to target locations (clovers) using eight keys of their keyboard associated with the movement rule of the Knight. For simplicity not all the 15x15 states are shown. Top-right: Points over trials. The solid line indicates the median and the shading the interquartile range. Bottom-right: Proportion of trials in the task that participants miss the target, that arrive sub-optimally or optimally. The height of the bars represents the median and the error bars the interquartile range.

to accomplish goals while still having uncertainty about what their actions do.

In the present work we developed a grid navigation task that allows the study of mapping learning and planning, and how they might interact to accomplish the task's goals. In our experiments participants moved a cursor from start to target locations using a rule based on the Knight chess piece. This mapping could be more closely comparable to the complexity of mappings humans acquire in real world scenarios, involving multiple actions with a non-trivial rule. Importantly, this contrasts with previous studies where either there is no mapping or the action space is intuitive or small (Kahn, Karuza, Vettel, & Bassett, 2018; Fermin et al., 2010, 2016).

Lastly, there has been a recent interest in the planning literature in designing behavioral tasks that better resemble the complexity of activities that humans perform (van Opheusden & Ma, 2019; van Opheusden et al., 2021; Schulz, Klenske, Bramley, & Speekenbrink, 2017), but that are also tractable to cognitive modelling (e.g., playing board games). We believe our task is a step in this direction, as the planning trajectories involved in our navigation task are non-linear given the Knight rule, and are also based on a mapping that is being learned in parallel.

Methods

Participants

Seventy five undergraduate students (33 males, 39 females, 3 non-binary and 1 preferred not to say; mean age = 19.7, sd = 1.7) from Princeton University were recruited through the Psychology Subject Pool. The experiments were approved by the Institutional Review Board (IRB). All participants provided informed consent before performing the experiment.

Apparatus and task design

All experiments were performed in person using the same computer equipment. Stimuli were displayed on a 60 Hz Dell monitor and computed by a Dell OptiPlex 7050'a machine (Dell, Round Rock, Texas) running Windows 10 (Microsoft Co., Redmond, Washington). Participants made their responses using a standard desktop keyboard or mouse. All experiments were run on the browser and hosted on Google Firebase. Subjects were seated in front of the computer and were asked to follow the instructions to begin the task.

In a 15x15 grid, participants had to move cursor in the form of a frog to target locations represented by clovers (Figure 1). The cursor could only move in directions determined by the movement rule of Knight chess piece. On a given position of the grid, the locations where the cursor could move to were shaded in orange. From a pilot study we found out that without this shading participants' performance was considerably poor. We believe the reason for this was because in addition to learning the mapping and planning with it, subjects had to figure out the scope of the mapping, which considerably increased the complexity of the task. Subjects moved the cursor using eight keys of their keyboard with the exception of one condition of Experiment 2 where they used their mouse (see below). The task consisted of five hundred trials with a time limit of two hours. Twenty different pairs of start-goal locations were presented throughout the task; the order was randomized and each pair showed up once before seeing all the pairs again. The cursor always appeared at the same starting location, and the targets were placed in grid states that were either one or three moves away from the cursor (see below). Importantly, only one pair of start-target locations appeared for a given trial. If participants arrived at the target using the minimum number of moves, they received one hundred points. Then, they would lose five points for every extra move. If they did not arrive at the target location in ten seconds, it was considered a miss, they received zero points and moved on to the next trial. A similar points system has been used in other grid navigation studies (Bera et al., 2021a).

Computational modeling

In order to explore the cognitive mechanisms that could give rise to participants' performance in our task, we assessed three computational models that incorporate mapping-learning and planning into the computation of action values as well as a persistence component aiming to capture habitual, repetitive behavior. We fitted these models to the data of Experiment 1.

Bayes + BFS model: In our first model, we assume that the learning of the mapping occurs using Bayesian updating. Previous work has shown that humans are able combine their past experiences and novel observations in a way that is consistent with this framework (Körding & Wolpert, 2006). In

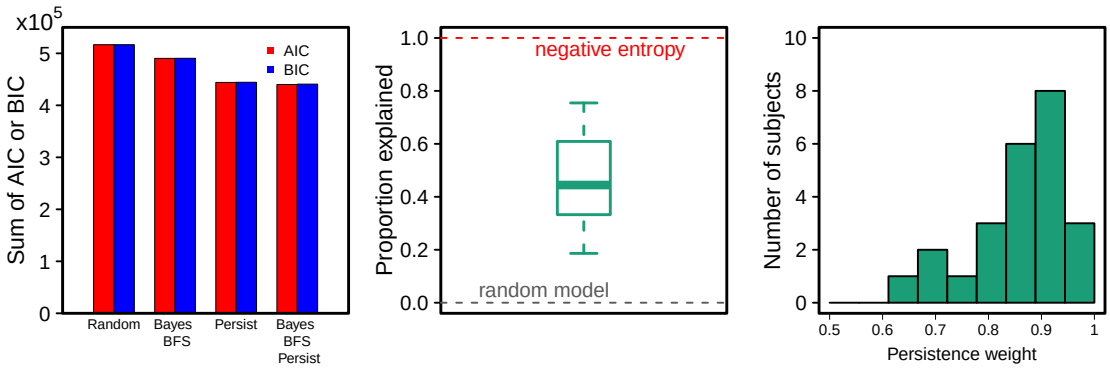

Figure 2: Modeling results from Experiment 1. Left: Sum of AIC and BIC across participants for each model. Lower values indicate a better fit. Middle: Proportion of the variability explained by the best fitting model (Bayes + BFS + Persistence) in our data based on the negative entropy and negative cross-entropy. Right: Persistence weight in our winning model.

particular, for every key, the cursor movement direction x is assumed to be generated by a Categorical distribution:

$$x_k \sim Cat(\theta_1, ..., \theta_8) \tag{1}$$

where $(\theta_1, ..., \theta_8)$ are the true probabilities that the key k moves the cursor to the eight possible directions of the Knight rule. These probabilities are unknown but can be inferred using Bayes rule. In order to do that, a prior distribution over $(\theta_1, ..., \theta_8)$ has to be specified which represents the initial knowledge of the mapping. For reasons of conjugacy, it is convenient to choose a Dirichlet distribution:

$$(\theta_1, ..., \theta_8) \sim Dir(1, ..., 1) \tag{2}$$

making the initial parameters equal to 1 gives no preference for any direction a priori. Then, the posterior belief about the mapping is described by another Dirichlet distribution:

$$(\theta_1^*, ..., \theta_8^*) \sim Dir(\alpha_1, ..., \alpha_8) \tag{3}$$

$$\alpha_n = 1 + \sum_{j=1}^{t} \mathbb{1}(j = i) \tag{4}$$

where $\sum_{j=1}^{t} \mathbb{1}(j = i)$ is the number of times the key was observed to go to the i direction up to trial t. The expected value of the parameters $(\theta_1^*, ..., \theta_8^*)$ can be computed to have a vector of probabilities instead of a vector of random variables:

$$\pi_i = \frac{\alpha_i}{\sum_{i=1}^{8} \alpha_i} \tag{5}$$

π_i is the probability that the cursor goes to the i direction. That is, if a key is pressed, the cursor can end up in the eight locations specified by the Knight rule with probabilities π. In model-based reinforcement learning π corresponds to the transition probabilities for a given state and action. Our model is a special case of these algorithms for which the tranistion probabilities are the same for all states. These probabilities are then used to compute the expected distance to the target in the next time step if that key was pressed:

$$E(d) = \sum_{j=1}^{8} d_i \pi_i \tag{6}$$

where d_i is the Knight distance to the target at the location the cursor would be if it moved to the i direction. In order to compute d, we used Breadth First Search (Erickson, 2019) on a Knight graph. In this structure, every node represents a grid state and nodes are connected among themselves if the cursor can reach them using the Knight rule. BFS is thought to represent the planning process in our model and Algorithm 1 shows the pseudocode to implement it. Briefly, what BFS does is to search on the Knight graph by first visiting the nodes that are one move away from the current location, then it checks if the target is there; if it isn't, then it continues searching in the nodes that are two moves away and so on. It continues this process until it reaches the target. We can then use $-E(d)$ to represent the value of pressing a given key. Changing the sign to negative makes lower distances more valuable.

Crucially, equation 6 formalizes the interaction between our learning (Bayes) and planning component (BFS): they are combined to obtain how valuable the actions are on a given state. The values $-E(d)$ for each key can then be plugged into a Softmax function to obtain the probability that key k is pressed at time step t :

$$\phi_t^k = \frac{e^{-\beta E(d)_t^k}}{\sum_{k=1}^{8} e^{-\beta E(d)_t^k}} \tag{7}$$

$$R_t \sim Cat(\phi_t^1, ..., \phi_t^8) \tag{8}$$

where β is the inverse temperature parameter and reflects decision noise. R_t is the response at time step t. This model has one free parameter: β.

Bayes + BFS + Persistence model: We considered a variation of our previous model which has a tendency to persist on responses that have been chosen in the past. We believe this is a plausible mechanism of choice when rewards are infrequent, and which could reflect habit formation. Similar mechanisms have previously been used in multi-arm bandits decision-making tasks (Miller, Botvinick, & Brody, 2021). In particular, the persistence component at time step t for key k is computed as follows:

$$P_t^k = P_{t-1}^k + \eta(r - P_{t-1}^k) \quad (9)$$

where $r = 1$ if a response for key k is generated and 0 otherwise. η is analogous to a learning rate parameter, here reflecting the weighting of recent responses. This algorithm is equivalent to a running average of the history of responses, where P_t^k will be higher the more responses have been given to key k in the past. P_t^k can be used to compute probabilities in generating responses as before:

$$\phi_t^k = \frac{e^{\beta P_t^k}}{\sum_{k=1}^{8} e^{\beta P_t^k}} \quad (10)$$

Then, equations 7 and 10 are linearly combined using the weight ω:

$$\hat{\phi}_t = \omega \phi_t^{Bayes+BFS} + (1-\omega)\phi_t^{Persistence} \quad (11)$$

$$R_t \sim Cat(\hat{\phi}_t^1, ..., \hat{\phi}_t^8) \quad (12)$$

This model has three free parameters: β, η and ω.

Persistence model: In order to evaluate the contribution of the persistence component on its own, we also considered equation 9 and 10 as a separate model. This model has two free parameters: η and β.

Random model: Finally, as a lower boundary, we evaluated a random model, which assigns equal probability to all responses at every time step:

$$R_t \sim Cat(\frac{1}{8}, ..., \frac{1}{8}) \quad (13)$$

Model fitting and model evaluation.

We fitted the proposed models to our data using Bayesian Adaptive Direct Search (Acerbi & Ma, 2017). Then we obtained the Akaike Information Criterion (Akaikei, 1973) and the Bayesian Information Criterion (Schwarz, 1978) to compare the models while penalizing for the number of parameters. Finally, we computed the negative entropy and the negative cross entropy as a measure of objective information content. The negative entropy represents the upper boundary of any probabilistic model (Shen & Ma, 2016). We estimated this quantity following the procedure described by Grassberger (Grassberger, 1988, 2003). The negative cross-entropy represents how much we can know about the data given an imperfect model. An estimator of the negative cross entropy is the logarithm of the likelihood function evaluated at the maximum likelihood estimates of the parameters, which we had already obtained when fitting our models. Based on the negative entropy and the negative cross-entropy we computed the proportion of the variability in our data that was explained by our best model.

Experiment 1

The goal of Experiment 1 (n=25) was to explore human performance on a task that necessitated simultaneous mapping-learning and planning. Additionally, we assessed whether their behavior can be captured by a model incorporating Bayesian updating, path search, and an optional persistence component, which could reflect learning, planning and habit formation. In the experiment, participants performed the grid navigation task shown in Figure 1, where the cursor was controlled using eight keys of the keyboard (A,S,D,F,H,J,K and L). Each key was associated with a move of the Knight. The particular direction each key moved the cursor to was randomized across participants. The target locations in this experiment were always three moves away from the starting location. It is important to notice that, since the movement directions were shaded in orange, participants didn't have to learn the rule itself, but its mapping to the keys.

Algorithm 1: BFS algorithm to compute Knight distance. *Note:*The function *Inside* returns whether the new location is inside the board.

Data: Size of the gird: M, N; Position of the cursor: X, Y; Position of the target: S, T.
Result: Knight distance to the target
int dx = [1, 2, 2, 1, -1, -2, -2, -1] ;
int dy = [2, 1, -1, -2, -2, -1, 1, 2] ;
Queue k;
$k.push([X,Y])$;
bool $visited[M+1][N+1] = False$;
$visited[X;Y] = True$;
int $moves[M+1][N+1] = 0$;
while $k \neq 0$ **do**
  z = $k.pop()$;
  if *z[0]==S and z[1] == T* **then**
    return moves[z[0]][z[1]];
  end
  for $i = 0; i < 8; i++$ **do**
    $new_x = z[0] + dx[i]$;
    $new_y = z[1] + dy[i]$;
    if
     $Inside(new_x, new_y, M, N) \&\& !visited[new_x][new_y]$
     then
      $visited[new_x][new_y] = True$;
      $moves[new_x][new_y] = moves[z[0]][z[1]] + 1$;
      $k.push([new_x][new_y])$;
    end
  end
end

Results

Behavioral Results: Given the complexity of the mapping and planning in our task, we did not have a priori expectations that subjects would be able to improve their performance. How-

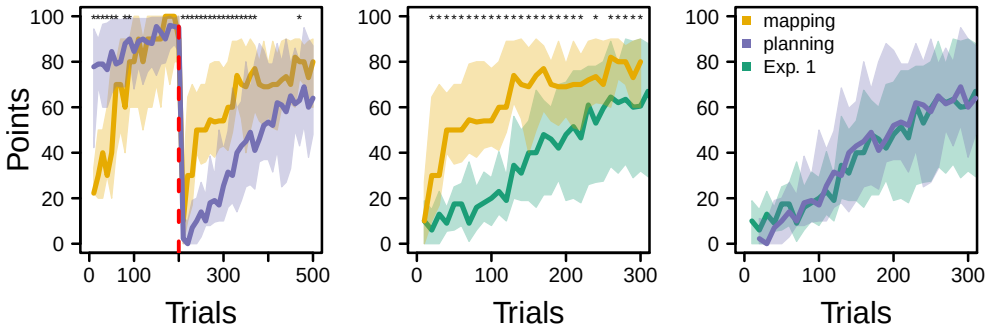

Figure 3: Behavioral results from Experiment 2. Left: Participants' points over trials for the mapping (gold) and planning (purple) conditions. The red line indicates the beginning of the full task as in Experiment 1. Middle: Participants' points during the full task for the mapping condition (last three hundred trials) compared to performance of subjects from Experiment 1. Right: Participants; performance in the full task for the planning condition compared to performance of subjects in Experiment 1. Solid lines represent the median and the shading the interquartile range. Asterisks at the top indicate significant differences between both group ($p < 0.5$).

ever, in the top-right panel of Figure 1, it is clear that we observed a canonical learning function. Performance appears to reach an asymptote at around 40-80 points, which means that subjects sometimes still did not arrive at the target or arrive suboptimally by the end of the experiment. In the bottom-right panel of Figure 1 we show the proportion of trials in the experiment that participants miss the target, arrive suboptimally (in more than three moves) or arrived optimally (in three moves). Overall, in more than seventy percent of the trials participants either miss the target or arrived suboptimally, which could reflect the complexity of the mapping and planning involved – a point we will explore in Experiment 2.

Modelling Results: In Figure 2 we show the modelling results. According to the sum of AIC and BIC across participants, the models with the persistence component (either on its own or with the Bayes + BFS components) clearly outperform the other models. Crucially, the persistence model on its own has lower performance than the Bayes + BFS + Persistence model (Δ summed AIC = $3.8x10^3$, Δ summed BIC = $3.5x10^3$) which suggests that the mapping-learning and planning components are important elements of subjects' choices. Additionally, in the middle panel, we note that there was a considerable variability in how good our best model (Bayes + BFS + Persistence) was able to explain our data, ranging from 18 to 75%. Finally, in the right panel we show the weight of the persistence component in our winning model. The high values of this parameter suggest that participants may have reduced the complexity of the task by repeating previous choices regardless of their optimality.

Experiment 2

In Experiment 1 we found that participants' improved their performance in our task though not always arriving optimally to the target. We also found that the mapping-learning and planning components were important elements of participants' choices given that, when included in a model, they outperformed a pure habit formation model. In Experiment 2 we asked whether breaking down the task into its mapping-learning and planning components would significantly improve performance. In order to do so, we ran two experimental conditions.

Mapping: In this condition (n=25), our goal was to test whether participants would improve their performance in our task by being exposed a priori to the mapping-learning component while removing the planning component. In order to do so, subjects performed the first two hundred trials of the experiment with target locations that were only one move away from the starting point. This way, a sequence of key presses would not need to be computed, as it would be the case for target locations being several moves away. In addition, participants were instructed to use those trials to learn the direction each key moved the cursor to. In the last three hundred trials, participants experienced the full task as in Experiment 1, with target locations being three moves away.

Planning: In this condition (n=25), we evaluated whether participants would improve their performance by being exposed only to the planning component of our task while removing the mapping-learning. In order to isolate planning, participants performed the task using their computer mouse for the first two hundred trials, rather than using their keyboard. As such, no novel movement-mapping had to be learned as participants could leverage their existing mapping of a computer mouse. Participants had to click on the locations of the grid where they wanted the cursor to move to, and the only valid locations were the ones shaded in orange according to the Knight rule. This way, they would not have the opportunity to experience the mapping, though they would have experience on how to arrive at the targets. In the last three hundred trials, participants were exposed to the full task as in Experiment 1.

Results

In Experiment 2, we tested whether prior experience with the mapping or planning component would benefit the sub-

sequent performance in our task. In the left panel of Figure 3 we observe that before the full task began (red line), the planning condition significantly outperformed the mapping one ($p < 0.05$) early in learning. However, this relation reversed by the time the full task began, where the mapping group had significantly better performance ($p < 0.05$) for around two hundred trials until both conditions reached around the same level of performance. Additionally, we compared the performance of these two conditions with Experiment 1, when participants were performing the full task. In the middle panel of Figure 3 we can see that training with the mapping before the full task provides a significant advantage over a group that started with the full task from scratch. Having experienced the planning component alone does not provide any benefit as compared to Experiment 1 (right panel of Figure 3), suggesting that the most efficient way to acquire a new skill is to learn the mapping before planning.

Discussion

When acquiring novel skills, humans face the double challenge of learning what the outcomes of their actions are and then how to use them sequentially to achieve future goals. The study of this matter lies at the heart of motor learning and planning research. However, both fields have addressed the topic mostly from separate fronts. On the one side, motor learning research has focused on sequence learning tasks where the planning component is constrained or limited by the experimenter. On the other side, planning studies rarely focus on situations where the action mapping is still uncertain or unknown. In the current work, we aimed to address this gap by exploring how mapping-learning and planning could interact in a task were both components are likely to be involved.

In Experiment 1, we found that participants were able to improve their performance over trials, although overall they did not choose optimal trajectories to the targets most of the time. This is potentially a consequence of the complexity of planning with a non-trivial action mapping, which may mirror real-life skills where optimal performance may not be achieved early on in training. Indeed, there has been renewed interest in developing experimental tasks that are sufficiently complex that capture the complexity of many scenarios that humans face in real life where there are multiple actions to choose from and multiple ways of combining them to achieve goals (van Opheusden & Ma, 2019; van Opheusden et al., 2021). Yet, at the same time that are still tractable enough for the use of relatively simple models that shed light on human cognitive mechanisms.

With this in mind, we note that our best model explained up to 75% of the variability in our data for some participants, but it was as low as 18 % for others, which leaves considerable room for improvement. In this model, the persistence component had a high influence in the overall output (Figure 2, right), which suggests that participants generated repetitive patterns of responses, potentially indicating the formation of a habit. Importantly, a pure persistence model performed considerably worse than our best model, suggesting that the mapping-learning and planning components are important elements of participants' choices in the task.

We have considered BFS as a starting point to represent planning as it is relatively simple to implement given its uninformed structure. This implies that no particular planning trajectory is favored in the search for the goal. However, this is not necessarily a realistic assumption in human planning. For example, in our task, trajectories going in the direction of the target might be favored with respect to trajectories going away from it. This simple scenario is not captured by BFS. Heuristic-based search algorithms like Best First Search can instead be used to specify preference over certain trajectories using a value function. For example, in a recent work van Opheusden and Ma (2021) used this algorithm to model how people plan in the 4-in-a-row game, pointing out that people might prune variations of the game that do not seem promising or intuitive, such as going away from the target in our task.

In Experiment 2 we explored whether breaking down the task into its mapping-learning and planning components would significantly improve performance as compared to Experiment 1. Indeed, we found that prior exposure to the mapping-learning, but not the planning component, provides a performance improvement in the full task. One explanation for this could be that the planning component was not as crucial to succeed in the task as we thought it would be. Therefore, having prior experience with it does not provide a significant advantage as compared to starting from scratch. Future work can test whether this hypothesis is true by making people find trajectories to the targets that are more involved, for example by adding obstacles or increasing the the distance to them. Remembering the trajectory solutions to those situations could turn out to be meaningful when the task has to be executed with a different controller (keys instead of mouse).

In addition, our results suggest that planning can be performed more efficiently when the action mapping has already been acquired. If both the mapping-learning and planning occur simultaneously since the beginning, performance can be lower as we observed in Experiment 1. This could shed light on the effectiveness of different cognitive strategies when acquiring a new skill. For example, we would be able to know why an amateur guitarist would improve slower if trying to play songs right from the begining instead of listening to the instructor's advice of practicing with scales first.

In summary, this work explores human performance in a behavioral tasks that involves learning a complex mapping and generating sequential decisions with it. We believe this type of experiments are a rich source of data to understand the interaction the cognitive mechanism that allow humans to acquire their vast repertoire of skills during their life.

References

Acerbi, L., & Ma, W. J. (2017). Practical bayesian optimiza-

tion for model fitting with bayesian adaptive direct search. *Advances in neural information processing systems*, *30*.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of experimental psychology: General*, *117*(3), 288.

Adams, J. A. (1971). A closed-loop theory of motor learning. *Journal of motor behavior*, *3*(2), 111–150.

Akaikei, H. (1973). Information theory and an extension of maximum likelihood principle. In *Proc. 2nd int. symp. on information theory* (pp. 267–281).

Balsters, J. H., & Ramnani, N. (2011). Cerebellar plasticity and the automation of first-order rules. *Journal of Neuroscience*, *31*(6), 2305–2312.

Bera, K., Shukla, A., & Bapi, R. S. (2021a). Cognitive and motor learning in internally-guided motor skills. *Frontiers in Psychology*, *12*, 604323.

Bera, K., Shukla, A., & Bapi, R. S. (2021b). Motor chunking in internally guided sequencing. *Brain Sciences*, *11*(3), 292.

Erickson, J. (2019). *Algorithms*. Independently Published.

Fermin, A., Yoshida, T., Ito, M., Yoshimoto, J., & Doya, K. (2010). Evidence for model-based action planning in a sequential finger movement task. *Journal of motor behavior*, *42*(6), 371–379.

Fermin, A., Yoshida, T., Yoshimoto, J., Ito, M., Tanaka, S. C., & Doya, K. (2016). Model-based action planning involves cortico-cerebellar and basal ganglia networks. *Scientific reports*, *6*(1), 1–14.

Fitts, P., & Posner, M. (1967). Human performance.

Grassberger, P. (1988). Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, *128*(6-7), 369–373.

Grassberger, P. (2003). Entropy estimates from insufficient samplings. *arXiv preprint physics/0307138*.

Hunt, L., Daw, N., Kaanders, P., MacIver, M., Mugan, U., Procyk, E., ... others (2021). Formalizing planning and information search in naturalistic decision-making. *Nature neuroscience*, *24*(8), 1051–1064.

Kahn, A. E., Karuza, E. A., Vettel, J. M., & Bassett, D. S. (2018). Network constraints on learnability of probabilistic motor sequences. *Nature human behaviour*, *2*(12), 936–947.

Kami, A., Meyer, G., Jezzard, P., Adams, M. M., Turner, R., & Ungerleider, L. G. (1995). Functional mri evidence for adult motor cortex plasticity during motor skill learning. *Nature*, *377*(6545), 155–158.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in cognitive sciences*, *10*(7), 319–326.

Korman, M., Raz, N., Flash, T., & Karni, A. (2003). Multiple shifts in the representation of a motor sequence during the acquisition of skilled performance. *Proceedings of the National Academy of Sciences*, *100*(21), 12492–12497.

Miller, K. J., Botvinick, M. M., & Brody, C. D. (2021). From predictive models to cognitive models: Separable behavioral processes underlying reward learning in the rat. *bioRxiv*, 461129.

Newell, K. M. (1985). Coordination, control and skill. In *Advances in psychology* (Vol. 27, pp. 295–317). Elsevier.

Newell, K. M. (1991). Motor skill acquisition. *Annual review of psychology*, *42*(1), 213–237.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., ... others (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609.

Schulz, E., Klenske, E. D., Bramley, N. R., & Speekenbrink, M. (2017). Strategic exploration in human adaptive control. *bioRxiv*, 110486.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Shen, S., & Ma, W. J. (2016). A detailed comparison of optimality and simplicity in perceptual decision making. *Psychological review*, *123*(4), 452.

van Opheusden, B., Galbiati, G., Kuperwajs, I., Bnaya, Z., Ma, W. J., et al. (2021). Revealing the impact of expertise on human planning with a two-player board game.

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