

**From the holdings of the Department of Special Collections , Princeton University Library,  
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

---

Doctoral Dissertations, 2011-2026

Psychology

---

2024

# BUILDING NOVEL ACTION-OUTCOME MAPPINGS FOR SEQUENTIAL MOTOR SKILLS

Velázquez Vargas, Carlos Alan

Princeton, NJ : Princeton University

---

<http://arks.princeton.edu/ark:/88435/dsp010g354j61m>

*U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>*

BUILDING NOVEL ACTION-OUTCOME MAPPINGS FOR SEQUENTIAL MOTOR SKILLS

Carlos Alan Velázquez Vargas

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY

THE DEPARTMENT OF PSYCHOLOGY

Advisers: Jordan A. Taylor, Nathaniel D. Daw

November 2024

© Copyright by Carlos Alan Velázquez Vargas, 2024. All rights reserved.

## Abstract

Understanding how humans acquire novel motor skills is a central topic in motor learning research. However, much of the work in this field has focused on adaptation experiments, leaving other key aspects of *de novo* skill acquisition less explored. For many *de novo* skills, individuals must learn new associations between discrete actions and arbitrary outcomes. This is evident in digital devices like video games, where pressing buttons on a controller can make a character jump or run. These action-outcome mappings are fundamental to the formation of the new skill. Therefore, understanding how they are learned and consolidated is essential for advancing our knowledge of motor skill acquisition and its application to various domains, from gaming to real-world tool use.

In Chapter 2, using a task of grid navigation, I study how these action-outcome mappings are acquired and examine the role of training variability in the formation of generalizable mappings. Crucially, when a novel mapping is being learned, it often occurs within the context of sequential decision-making, allowing the interaction of motor learning and planning. In Chapter 3, I investigate this interaction with the aim of bridging the gap between motor sequence learning and planning research. Finally, in Chapter 4, I study the effectiveness of external contextual cues in the learning of multiple mappings, which have proven unsuccessful in standard motor adaptation experiments. The behavioral results from each chapter of this dissertation are complemented by computational models that integrate algorithms from reinforcement learning, tree search, and Bayesian learning. These models aim to provide insights into the cognitive processes underlying participants' performance.

## Table of contents

|                                                                                            |     |
|--------------------------------------------------------------------------------------------|-----|
| Acknowledgments.....                                                                       | 5   |
| Chapter 1: Introduction.....                                                               | 7   |
| Chapter 2: Model-free and model-based contributions for novel action-outcome mappings..... | 13  |
| Chapter 3: Planning with novel action-outcome mappings.....                                | 79  |
| Chapter 4: Separating representations of distinct action-outcome mappings.....             | 110 |
| Chapter 5: General conclusion.....                                                         | 125 |
| References.....                                                                            | 133 |
| Publications.....                                                                          | 148 |

## Acknowledgments

I can still recall the feeling of seeing Princeton for the first time. It was like a dream. The beauty of the campus and the academic atmosphere filled me with hope and energy for the new chapter of my life that was about to begin. I remember part of the speech from the graduate school dean in the welcoming session “...And keep in mind, this is going to be a marathon for you, so do not rush”. Five years later, I truly understand the meaning of those words. It was indeed a long race—and most importantly, one I could not have finished alone. Throughout these five years, I have been surrounded by my beloved family and friends and I would like to acknowledge and thank them for making this journey possible:

To my mom and dad. For always believing in me. For teaching me to keep moving forward and to enjoy what I do in life. For their ceaseless effort and love to continue supporting me before and during the PhD. For keeping my motivation up and for always listening to me. To my brother, for bringing a smile every time I see him, and for taking care of me when I need him the most. To my grandparents for reminding me the value of hard work and resilience.

To my dear friends in México — Victor, Alondra, Casandra, Zaori, Lili, Aurora, Itzel, Gustavo, Diana, Uriel — for loving me as I am and for keeping me strong, even from afar. I miss you all.

To the great friends I made during these years: Thiago, Isaac, Felix, Jonathan, Rachel, Sev, Cameron, Clotaire, Yihang. For making Princeton an enjoyable journey, and more than just an academic place.

To Jordan, for his patience, for showing me that a supervisor can also be a friend. For his support and encouragement during the hard times of COVID. For pushing me to achieve my best academically and for the many intellectual and casual conversations —you really made Princeton a special and fun place.

To Arturo Bouzas, for inspiring me to become a psychologist. For his invaluable mentorship all these years. For always keeping a good sense of humor and a relentless dedication to improving psychology and education in México.

To Michael Lee and Robert Wilson, for their mentorship and support that allowed me to get to Princeton.

To Victor, for his guidance since secondary school. For teaching me the love for mathematics and for becoming my cherished friend. For the many insightful conversations and advice. For the great trips, food and drinks along the way!

To Alejandra, my psychologist, for her wisdom and guidance during difficult times. Thank you for helping me find meaning and purpose in what I do.

To the current and past members of the IPA lab that I met: Jonathan, Olivia, Chandra, Sarah, Yiyu, Tommy, Yifei, Mohan.

To Nathaniel and Casey for their helpful and insightful feedback along the program. To Natalia and Yael for kindly accepting being part of my oral committee.

Finally, I would like to thank the Department of Psychology at Princeton University for making a great effort to build an inclusive and supportive community.

# Chapter 1

## Introduction

During their lifetime, humans can develop a wide and complex repertoire of motor 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. First, humans must 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 learning of these mappings poses the first challenge when attempting to acquire a novel motor skill (Fitts & Posner, 1967; Adams, 1971; Ackerman, 1988; Newell, 1985, 1991). 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; Mattar and Lengyel, 2022). In this dissertation, I aim to further understand how complex motor skills—requiring both sequential decision-making and the learning of a novel mapping—are acquired. Additionally, I aim to shed light into how the vast repertoire of such skills is represented while avoiding catastrophic interference.

Traditionally, the formation of motor mappings has been studied in the context of motor adaptation, where humans adjust their performance in response to environmental perturbations. For example, a professional golfer may need to re-aim their shot to compensate for wind direction. Similarly, in heavy snow, an individual may learn to adjust their leg movements by lifting them higher and moving more slowly to avoid falling. While new mappings resulting from adaptation are crucial for maintaining motor skill performance, many of the skills that humans

acquire require the formation of motor mappings *de novo*. For example, when learning to play the guitar, particularly if no other instrument has been played before, the link between hand movements and the desired outcomes (e.g., notes or chords) must be learned presumably from scratch. Although both adapted and *de novo* mappings are important for successful motor performance, throughout this dissertation I focus exclusively on the latter.

To do so, the experiments presented in the forthcoming chapters leverage a grid navigation task originally developed by Fermin et al., (2010), which has recently been adapted in other studies (Franklin and Frank, 2020; Bera et al., 2021; McDougale et al., 2022). Here, participants are tasked to move a cursor from start to target locations in a grid world using their keyboard keys. Each key moves the cursor in arbitrary directions which participants need to learn in order to navigate the cursor to the goal location using a trajectory (sequence) that minimizes the number of key presses. A key distinction of this task compared to previous sequence learning experiments is that participants are free to choose the sequences to reach the targets, rather than being prespecified by the experimenter. These, *internally generated sequences* (Bera et al., 2021), are likely to underlie improvisation in activities such as sports or playing musical instruments, where individuals need to choose the best course of action, without instruction, to achieve a goal.

In Chapter 2, I hypothesize that the quality of the learned representation (i.e., mapping) can be distinguished by their generalization power. For example, mappings based on local state-action associations can be learned in situations where task variability is low, allowing the formation of habits at a low computational cost. However, they can result in poor generalization. On the other hand, function-based mappings, which link responses to global state transitions, can be trained in situations with higher variability. This approach allows for generalization and planning to novel goals but requires a higher computational cost. In order to manipulate the training variability, participants are exposed to different numbers of start-target locations in the

grid navigation task. Then, they are tested on generalization to novel targets, aiming to reveal the acquired mapping. Additionally, to gain insight into the cognitive mechanisms underlying these distinct mappings, I evaluate computational models that formalize them and compare the models' performance with empirical data from four experiments.

In Chapter 3, I investigate how more complex mappings, involving multiple actions with non-trivial transitions, are learned and used for planning. The goal is to better understand real life motor skills which can take hours or even days to be fully acquired. In order to do so, participants perform the grid navigation task using a mapping based on the piece of chess known as the Knight. The Knight has eight possible moving directions which follow L-shape transitions. In two experimental studies, I address how separating the “mapping learning” component from the “planning” component of this task, can enhance performance. In addition, I assess participants' performance for different horizons (i.e., target distances) seeking key signatures of planning such as the increase in reaction time with the planning depth. The results of these experiments are interpreted using computational models that incorporate both motor learning and a planning process with a limited depth, which aims to capture humans' limited cognitive resources.

Lastly, it is well-known that humans can learn multiple motor mappings, for example, when playing different video-games or musical instruments, raising the question of how we separate the representations and avoid catastrophic interference? To answer this question, in Chapter 4, I developed a variation of the grid navigation task where subjects have to learn two mappings which are randomly interleaved over trials. Particularly, I ask whether contextual cues implemented as different “grid worlds”, allow participants to learn both mappings successfully. While similar contextual cues are not effective in preventing interference in standard motor adaptation tasks (Howard et al., 2013), I ask whether they can prove successful in *de novo* skill learning, as it is known to involve different brain mechanisms (McDougle, 2022). When a skill is

acquired *de novo*, new control policies are created, rather than recalibrating existing ones (Krakauer et al., 2019.). In addition, I evaluated the performance of two meta-reinforcement learning agents in the same task and analyzed their internal representations to understand how humans could be representing the task as well.

Together, the studies in this dissertation aim to explore how humans acquire motor skills involving discrete and internally generated sequences. Given the prevalence of these skills in various activities involving digital devices, the insights gained from these studies can enhance our understanding of such skills and inform better strategies for their acquisition. Additionally, this research can lay the groundwork for studying skills that develop in continuous domains, such as tool manipulation or learning to play some musical instruments.

## Chapter 2

### Introduction

The first problem to be overcome in learning any novel motor skill is to associate particular actions with desired outcomes. This problem has become increasingly complex in the digital age, where the mapping between actions and outcomes can be as diverse as the imagination allows – just consider the variety of action-outcome associations underlying digital applications and video games. For example, using two thumbs to type a text message, using a pinch motion to zoom in and out of content on a smartphone, or steering a car in a video game. At first, learning these novel mappings is cumbersome and effortful but as learning progresses a mapping between actions and outcomes is eventually formed, allowing the individual to use the device successfully with ease. The formation of this mapping is arguably one of the most important steps for learning any new motor skill (Fitts & Posner, 1967; Adams, 1971; Ackerman, 1988; Newell, 1985, 1991). Surprisingly, however, we know very little, with a few exceptions (Mosier et al., 2005; Liu et al., 2011), about how novel motor mappings are initially formed.

Traditionally, the question of how motor mappings are learned has been the focus of sensorimotor adaptation tasks (e.g., prisms, visuomotor rotations, and force fields), which impose a perturbation on the sensory outcome of a movement (Shadmehr and Mussa-Ivaldi, 1994; Martin et al., 1996; Krakauer et al., 2000). While adaptation tasks were originally thought to serve as a model paradigm to study this question (Jordan et al., 1992; Miall and Wolpert, 1996; Shadmehr et al., 2010), in recent years, it has become clear that these tasks may only pressure the recalibration of an existing motor mapping when faced with an externally imposed perturbation – not the establishment of the mapping in the first place (Krakauer et al., 2019; Hadjiosif et al., 2021). Only when these recalibration mechanisms fail to fully counteract these perturbations in adaptation paradigms (Hadjiosif et al., 2021), may more *de novo* learning engage to develop a new controller for the task (Yang et al., 2021).

While there have been numerous studies of operant conditioning and associative learning, linking actions to outcomes, it is unclear the degree to which learning in these studies reflects the formation of a motor mapping *per se* (Thorndike, 1927; Petrides, 1997; Elsner and Hommel, 2004). It can be helpful, at least conceptually, to distinguish two levels of choice: a more abstract, internal level of reasoning about goals and state changes, and a more external, response-focused level about how to use movements to bring these plans to fruition. In many studies, such as in spatial or maze navigation, the agent already knows the control policy of how to move (i.e., how an action leads to a state change) and instead the focus is on reasoning or learning at a more abstract level how the state change leads to a desired outcome in terms of reward (Sutton and Barto, 1998; Simon and Daw, 2011). Conversely a different set of paradigms focus entirely on externally-cued responses, without any internal plan. Such tasks include motor sequence learning (Nissen and Bullemer, 1987; Willingham et al., 1989; Curran and Keele, 1993), discrete sequence production (Verwey, 2001; Abrahamse et al., 2013), and  $m \times n$  tasks (Hikosaka et al., 1995; Bapi et al., 2000), all of which can be viewed as a form of *de novo* motor learning, establishing a relationship between arbitrary actions and outcomes. However in these studies, there is no underlying mapping from internal goals, from which a generalizable, motor map may form. Generalization to new situations or contexts is considered a hallmark feature of a motor mapping, as opposed to rote memorization of stimulus-response associations (Shadmehr and Mussa-Ivaldi, 1994; Mussa-Ivaldi, 1999; Shadmehr, 2004).

Furthermore, the different variants of these sequence learning tasks are externally generated, such that the appropriate sequence of responses is fully specified by the experimental stimuli (Bera et al., 2021). The participant must precisely follow the set of stimulus-response pairs to be successful in the task. As such, they may only reflect a subset of the kinds of motor skills that we perform in everyday life that are internally generated. Thus, while there has been tremendous progress in understanding how externally-generated,

stimulus-response mappings are learned, there has been comparatively less progress in understanding how internally-generated, response-outcome mappings are formed.

The increased complexity and degrees of freedom available when learning an internally-generated, response-mapping may be one potential reason why progress has been slow. The difficulty is in designing and studying a task that is in the “Goldilocks zone” between sufficient experimental complexity and analytical tractability (Opheusden and Ma, 2019). Fermin and colleagues developed the *grid navigation task* to fit within this zone to study the core problem of learning an internally-generated mapping: The formation of a novel and arbitrary motor mapping (Fermin et al., 2010, 2016). Here, participants learn to navigate a cursor from various starting locations to various target locations on a grid through a series of keypresses. The goal is to navigate to the target location in the minimum number of moves as quickly as possible. Importantly, there is an unintuitive and arbitrary mapping between the keys and cursor movement that must be learned.

While in simple versions of this task participants can learn within a relatively short period, it remains an open question as to how this is accomplished (Fermin et al., 2010; Bera et al., 2021; Dundon, 2021). The formation of a new mapping is not always guaranteed. If the task only demands the repetition of a limited set of actions, then only local state-action associations may be learned – a form of rote memorization, which is likely what occurs in studies of sequence learning. However, if there is a greater degree of variability in training, then a richer representation of skill may be learned, such as the formation of an internal model between the action-outcome space. This would afford the ability to generalize outside the range of training (Schmidt 1975; Newell and Shaprio, 1976; McCracken and Stelmach, 1977; Kerr and Booth, 1978; Catalano and Kleiner, 1984; Berniker et al., 2014) – an idea that echoes classic theories of stimulus variability in learning (Estes and Burke, 1953; Raviv et al., 2022). These two forms of learning mirror the instance-based and algorithmic processes of a classic theory for automatization (Logan, 1988), as well as the more modern notions of model-free and

model-based reinforcement learning (Sutton and Barto, 1998; Daw et al., 2005, 2011; Haith and Krakauer, 2013; Raviv et al., 2023). In particular, the latter formalism seems well suited to capturing the candidate mechanisms. Model-based reinforcement learning is well suited to capture the covert formation of an abstract, internal plan that can then be generalizably realized through a separately learned motor mapping. This leads to the hypothesis that, much as in other circumstances such as operant leverpressing (Daw et al., 2005), simpler model-free (stimulus-response) learning will instead dominate when a narrow range of actions is overtrained.

Here, through a series of experiments, we seek to build on this work by leveraging the grid navigation task as a model paradigm to study how novel and arbitrary motor mappings are initially formed and seek to characterize how they may be learned through the model-free and model-based reinforcement learning framework. We hypothesize that the formation and representation of a novel motor mapping depends on the particular conditions of training. Specifically, the degree of exploration between the number of potential action goals and possible solutions to achieve that goal may pressure formation of a generalizable motor mapping over local state-action associations (e.g., rote memorization of specific sequences of actions). Generalization to untrained conditions will provide a key test for the existence of a motor mapping.

## **Materials and Methods**

**Participants:** 112 undergraduate students (49 males, 58 females, 4 non-binary and 1 preferred not to say; mean age = 19.9, sd = 1.4) from Princeton University were recruited through the Psychology Subject Pool. Sample sizes were based on prior studies of the grid navigation task (Fermin et al, 2010, 2016; Bera et al., 2021). The experiments were approved by the Institutional Review Board (IRB) and 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. All experiments were programmed in CSS, Javascript and HTML, and run on a web 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.

We employed a variant of the grid navigation task based on Fermin et al. (2010) in which participants were required to navigate a cursor from a starting position to a target location on a 9x9 grid using the J, K and L keys of their keyboard. On Experiment 1-3, each key moved the ship deterministically to one of three possible directions: right, down-left or up-left. On Experiment 4, the keys' directions followed a stochastic rule (see below). At the onset of the experiment, participants were provided with the following instructions *"In this game, you will use the letters J, K and L of your keyboard to move a vehicle through a grid to a target location. Your goal is to arrive using the shortest route. If you arrive with the shortest route, you will see a happy face. If you arrive using a different route, you will see a neutral face. If you do not arrive after a certain time, you will see a sad face."* After participants confirmed they understood the instructions, the task began. The cursor was displayed as a ship and the targets as anchors. Additionally, to make the task more engaging, it was performed with a background of the ocean with quiet wave sounds and 'bubble' sounds every time the cursor moved.

On a given trial, the cursor and a target appeared in locations that varied across experiments (see Figure 2.1). Depending on the performance of the trial, subjects could receive three types of feedback. If they did not arrive at the target in less than 10s, a sad face appeared in place of the target, along with a "wrong sound" indicating they had failed. If participants arrived at the target but not in the minimum number of key presses, a neutral face and sound were presented. If they arrived using the minimum number of key presses, a happy face with a

“correct sound” was presented. The visual feedback remained on the screen for 1s after which an inter trial interval of 500 ms occurred. Then, the next trial began. The experiment was divided into a training and a generalization phase, which will be described in detail for each experiment below. During the training phase of all experiments, the targets were placed seven moves away from the start location.

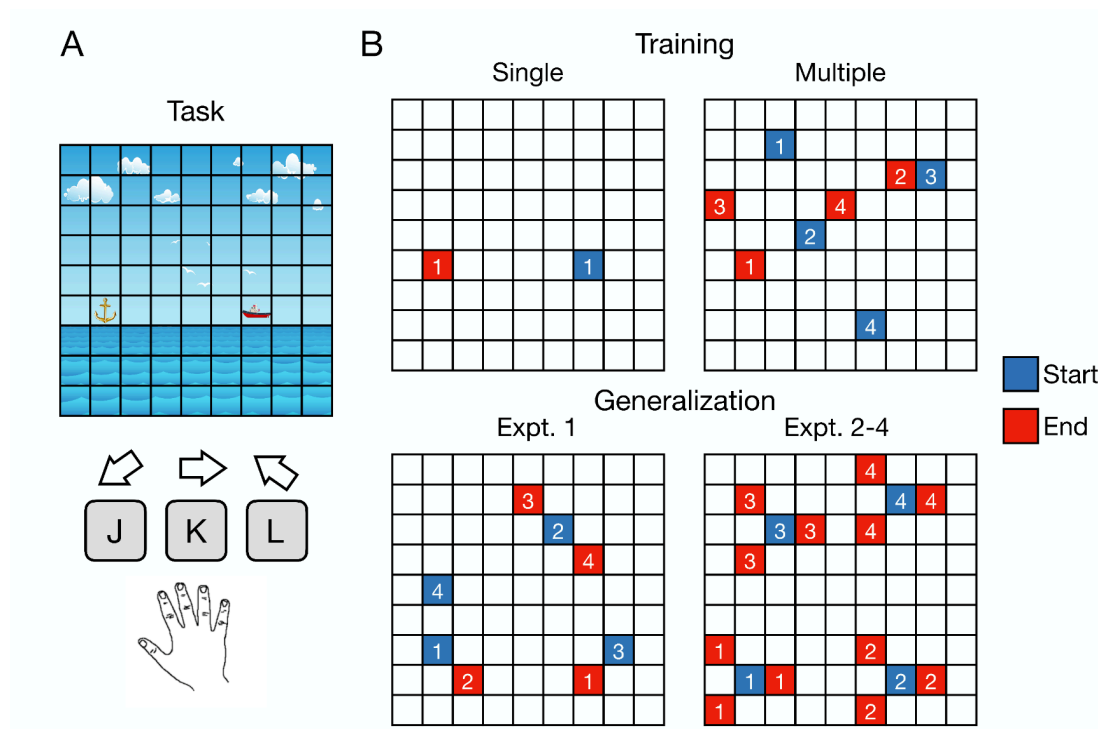


Figure 2.1: Experimental task. **(A)** Participants moved a cursor (ship) from start to target (anchor) locations in a grid environment. For Experiments 1-3, participants used a deterministic visuomotor mapping of three keys with moving directions: bottom-left, right and top-left. In Experiment 4, the mapping randomly changed after each keypress with a probability of 0.2 to any of the remaining directions. **(B)** In the four experiments, participants were trained with a single or multiple start-target pairs (see *Materials and Methods* for details). A generalization phase was presented after training where the target locations were either seven (Experiment 1) or one move away (Experiments 2-4) from the starting point. Blue and red grid states represent the potential start and target locations where the numbers indicate a specific pair. Only one pair was presented per trial.

**Experiment 1 Procedure:** The goal of this experiment was to determine whether specific training regimes with different levels of variability promote the formation of local state-action associations or flexible and generalizable visuomotor mappings. For all participants, the J, K and L keys moved the ship to the down-left, right and up-left, respectively (Figure 2.1A). The training phase consisted of 260 trials which were followed by 20 generalization trials interleaved with 20 training trials, giving a total of 300 trials. Subjects were randomly assigned to one of two groups that differed in the number of start-target pair locations presented during training (Figure 2.1). In the Single group (n=16), a unique start-target pair was presented for all training trials. In this case, the target could be reached using a unique sequence of key presses (e.g., J-L-J-L-J-L-K), however, participants were not constrained or encouraged to do so.

For the Multiple group (n=16), four start-target pairs were presented throughout training, where each of them appeared 65 times. We randomized the pairs such that the same pair did not show up more than twice in a row and all four pairs appeared once before observing them again. Additionally, the target for each pair could not be reached using the same sequence of key presses that arrived at other targets. In the generalization trials, four novel start-target pairs were presented for both groups. The target was placed seven moves away from the starting point just as in the training trials. Each of the generalization pairs was repeated five times but no pair appeared more than twice in a row, and all four pairs were observed before showing them again. No performance feedback (emoji faces and sounds) was provided in generalization trials, but movement feedback after each move was available. However, movement feedback was provided in both the interleaved training trials and the generalization trials.

**Experiment 2 Procedure:** In Experiment 2, we tested whether differences in generalization performance between the Single and Multiple groups in Experiment 1 would remain when movement-related feedback was withheld and no sequential planning was required. To achieve this, we modified the experimental setup by situating the target locations

during generalization trials only one step away from the starting point (as opposed to seven in Experiment 1). Moreover, we removed the training trials interleaved in the generalization phase.

The novel start-target pairs for generalization were created by linking four start locations with three possible target locations (Figure 2.1B). All pairs were presented at least once, and the remaining generalization trials were randomly chosen without replacement from the available twelve. Finally, In order to control for mapping-specific effects, we randomized the directions each key was assigned to across subjects. The training phase was the same as in Experiment 1 for the Single ( $n=16$ ) and Multiple ( $n=16$ ) groups. The total number of trials was 280.

**Experiment 3 Procedure:** In Experiment 3, we tested whether a short exposure to the training variability followed by a long exposure to no variability would be sufficient to learn the mapping as well as to maintain a memory of the mapping, thus affording good performance in the generalization phase at the end of the experiment. Participants ( $n=16$ ) first trained during 80 trials with four start-target pairs (Multiple trials), then experienced a first generalization phase of 20 trials with target locations being one move away as in Experiment 2. We chose 80 training trials of training given that in Experiment 1 and 2, asymptotic performance was reached in this time frame by the Multiple group. In addition, in online pilot study, we found that successful generalization was possible with this number of trials in the Multiple group. Following the first generalization phase, participants were exposed to 1000 trials with a single start-target pair (i.e., Single trials), which was the same as in the Single group of Experiment 1 and Experiment 2. The number of trials in this stage was decided based on a pilot study where we first tested 400 Single trials and found no decline in performance in generalization. Therefore, we aimed to extend this phase further to more robustly test the effect of early training with variability. Finally, participants were exposed to a second generalization phase of 20 trials. Importantly, the order of the start-target pairs in the two generalization phases was randomized.

**Experiment 4 Procedure:** The goal of Experiment 4 was twofold. First, we sought to test whether the visuomotor mapping was represented explicitly or implicitly. Second, we tested whether training variability induced by stochasticity in the key-to-direction mapping during training would prevent explicit memorization of the sequence and pressure learning of the mapping, which would afford generalization. To accomplish this, we imposed a probabilistic rule over the movement of the cursor. Specifically, during the training phase of both the Single (n=16) and Multiple (n=16) groups, there was a 0.2 probability that, on every move, the key moved the cursor to any of the other seven directions different from the original mapping (left-down, right and left-up). We use this probability based on previous studies on sequence learning that have found that adding this level of stochasticity prevents participants from learning sequences explicitly (Jiménez et al., 2006; Schvaneveldt and Gomez, 1998). In order to evaluate subjects' awareness of the visuomotor mapping of the task, we asked them at the end of the experiment to indicate the direction each key moved the cursor to. Specifically, pictures of the keyboard keys were displayed on the screen (J, K and L), each of them followed by eight moving options indicated with arrows (top, top-right, right, down-right, down, down-left, left and up-left). Participants had to select among the options the one they believed was the true moving direction of the key. Generalization trials were the same as in Experiment 2, with targets being one step away from the start locations and no feedback was provided.

**Behavioral data analysis:** All analyses were performed using the R statistical software (R Core team, 2023) or Matlab version 2022a (Natick, MA: The Math Works, Inc., 2022). Our main behavioral measure was optimal arrivals and reaction times to the target in Experiments 1-3, which was defined as the minimum number of key presses to move the cursor to the target (7 moves or 1 move, depending on the experiment and phase). In Experiment 4, due to stochasticity, our primary measure was simply arrival to the target even if it was not in the minimum number of key presses. We also examined both reaction time, defined as the time between target presentation and the first key press, and inter-keypress interval, defined as the

time between each successive key press. These behavioral metrics were binned every ten trials. When relevant comparisons were done between our Single and Multiple groups, we used Welch's t tests for unequal variances (Welch, 1947). Paired t tests were used for comparisons in Experiment 3, as samples were dependent. When comparing the learning curves between the Single and Multiple groups in Experiments 1, 2 and 4, we performed a mixed effect model with time and experimental group as fixed effects, and with subjects as random effects.

**Computational modeling:** In order to gain mechanistic insight into the learning processes that could have given rise to the results of our experiments, we evaluated five computational models in all our experiments. At one end of the modeling spectrum, we implemented a prediction error RL model to characterize inflexible, habitual behavior, which we believe could be induced in our Single group (model-free). Although this model works in a relatively straightforward manner, it predicts poor generalization as it can only know what to do in situations it has experienced in the past (Figure 2.4). At the other end of the modeling spectrum, we used a Bayesian model along with a tree-search planning process, to represent a learner that acquires the true key-to-direction mapping and leverages it to decide the best course of action (model-based). This mapping can also be thought to capture the function of the keys, and therefore, I also refer to it as function-based mapping. As we will describe below, this model would be able to generalize well in our task and we believe a similar mechanism could be drawn on in our Multiple group. Finally we considered three hybrid models that differ in the specification of the mixing weight between the model-free and model-based algorithms. In particular, we considered a hybrid model with a single weight, with two weights (one for the training phase and one for the generalization phase) and with a time varying weight where the arbitration between the systems is based on the familiarity of the current state.

*Model-free (MF):* This model uses prediction errors to update the value of the keys at each grid cell for every target location using absolute coordinates. Model-free algorithms have received considerable attention in the past years due to its simple trial-and-error mechanism,

which can capture a wide variety of behavioral and neural data (Schultz et al., 1997; Sutton & Barto, 1998; Miller et al. , 1995). In our task, it updates the value  $v$  for pressing key  $k$  after a prediction error  $\delta$  is observed. More explicitly, for every time step  $t$  :

$$v_t^k = v_{t-1}^k + \alpha \delta_{t-1}^k$$

$$\delta_{t-1}^k = r_{t-1}^k - v_{t-1}^k \quad ,$$

where  $r$  is the reward obtained and  $\alpha$  is a free parameter that modulates the speed of learning. We define reward  $r$  in terms of the reduction of the chessboard distance  $d$  to the target. Specifically:

$$r = 1 \quad \text{if } d_t < d_{t-1} \quad \text{or} \quad r = 0 \quad \text{if } d_t \geq d_{t-1}$$

$$d = \max(x_{\text{target}} - x_{\text{cursor}}, y_{\text{target}} - y_{\text{cursor}}),$$

$x$  and  $y$  are the grid coordinates of the target and cursor. Then, the probability for pressing key  $k$  at time step  $t$  is generated using a Softmax function:

$$\phi_t^k = \frac{e^{\beta v_t^k}}{\sum_{k=1}^3 e^{\beta v_t^k}}$$

$$R_t \sim \text{Cat}(\phi_t^1, \phi_t^2, \phi_t^3) \quad ,$$

where  $\beta$  is the inverse temperature parameter and  $R_t$  is the key press at time step  $t$ . This model has two free parameters:  $\alpha$  and  $\beta$ . Note that whereas many model-free approaches (temporal-difference methods, etc.) to multi-step decision tasks of this sort recursively learn a multi-step value function measuring distance to goal, here we streamline this approach slightly by defining the target value at each step non-recursively, in terms of the simple chessboard heuristic at each step. This is similar to advantage learning (itself a variant of the actor-critic), but with the value function component fixed as the chessboard distance. In a supplementary analysis (Figure S2.10), we show that this model-free algorithm performs better or at the same level than SARSA, a commonly used temporal-difference algorithm (Sutton and Barto, 1998).

We believe that the reduction in the chessboard distance is an intuitive measure of reward in this model, as it is equivalent to visually getting closer to the target. In addition, given the relatively few explicit rewards in our task (emoji faces at the end of the trial) compared to the number of moves, incorporating the distance reduction as a reward signal aimed to aid learning in the model to more closely resemble the fast learning curve observed in the data. However, this form of distance assumes the cursor can move to any of the adjacent locations, which is not true in our experiments, but is reasonable in an agent that has no knowledge of the key-outcome mapping. As we will see in our next model, the distance to the target can instead be measured as the number of key presses away from it. When the available moves of the cursor are constrained, the key-press distance can differ from the chessboard distance. More importantly, knowing the key-press distance implies knowledge of the true key-outcome mapping, a fundamental property of our next model.

*Model-based (MB)*: In this model, a probability distribution over the key-outcome mapping is updated using Bayes rule and subsequently used to reduce the number of key presses away from the target. In particular, for every key  $k$ , the cursor movement direction  $x$  is assumed to be generated by a Categorical distribution:

$$x_k \sim \text{Cat}(\theta_1, \dots, \theta_8)$$

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

$$(\theta_1, \dots, \theta_8) \sim \text{Dir}(1, \dots, 1).$$

Making the initial parameters equal to 1 gives no preference for any direction *a priori*. While we believe this is a reasonable starting point, the specification of the initial values of the

parameters can affect the sensitivity of the model to the data. For example, larger values can make the model less sensitive to data. Then, the posterior belief about the mapping is described by another Dirichlet distribution:

$$(\theta_1^*, \dots, \theta_8^*) \sim \text{Dir}(\alpha_1, \dots, \alpha_8)$$

$$\alpha_n = 1 + \sum_j \mathbb{1}(j = i),$$

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

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

$\pi_i$  is the probability that the cursor goes to the  $i$ th direction. That is, if a key is pressed, the cursor can end up in the eight adjacent locations with probabilities  $\pi$ . In model-based reinforcement learning  $\pi$  corresponds to the transition probabilities for a given state and action. Our model is a special case of these algorithms for which the transition 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:

$$E(d) = \sum_{i=1}^8 d_i \pi_i,$$

where  $d$  is the actual distance to the target, that is, the number of key presses away from it. In order to compute  $d$ , we used Breadth First Search (BFS; Moore, 1959). BFS transforms our grid environment into a graph where each node represents a grid state and nodes are connected among themselves according to the possible transitions in the grid given the visuomotor mapping. BFS is thought to represent the planning process in the model-based algorithm which

is known to work well in deterministic environments like ours (Mattar and Lengyel, 2022). What BFS does is to search on the graph created with the grid environment 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 use  $-E(d)$  to represent the value of pressing a given key. Changing the sign to negative makes lower distances more valuable, then these quantities can be plugged into a Softmax function:

$$\phi_t^k = \frac{e^{-\beta E(d)_k}}{\sum_{k=1}^3 e^{-\beta E(d)_k}}$$

$$R_t \sim \text{Cat}(\phi_t^1, \phi_t^2, \phi_t^3).$$

This model has one free parameter:  $\beta$ . Importantly, this algorithm has a different reward signal (the actual distance to the target) compared to the model-free algorithm (chessboard distance). While this is not a common assumption in model-free and model-based models, we reasoned that it could capture the idea that the model-free algorithm does not have access to values resulting from using a model of the world (transition probabilities), whereas the model-based algorithm does.

*Hybrid models:* As noted previously (Daw et al., 2011; Gläscher et al., 2010; Kool, 2017; Lee et al., 2014) we considered the possibility that participants implemented both model-free and model-based computations. We represented this possibility as weighted sums between the outputs of the model-free and model-based algorithms.

*Single-weight model (1W):* In this hybrid model, there is a single weight across the entire experiment. In particular:

$$\phi_t^{\text{Hybrid}} = \omega \phi_t^{\text{MB}} + (1 - \omega) \phi_t^{\text{MF}}$$

$$R_t \sim \text{Cat}(\phi_t^{\text{Hybrid} [1]}, \phi_t^{\text{Hybrid} [2]}, \phi_t^{\text{Hybrid} [3]}),$$

where  $\omega$  is the weight for the model-based component. Single-weight hybrid models provide a simple way to specify the interaction between the learning systems (Daw et al., 2011; Gläscher et al., 2010), although they could miss important information about the dynamics – if existent at all. This model has four free parameters:  $\alpha$ ,  $\beta_{MF}$ ,  $\beta_{MB}$  and one weight parameter  $\omega$ .

*Two and four-weight models (2W and 4W):* These hybrid models consider the possibility that the model-based weight,  $\omega$ , is different during the learning and generalization phases. Such variation can incorporate scenarios where the model-based influence increases when the novel targets are experienced. The 2W hybrid model has five free parameters:  $\alpha$ ,  $\beta_{MF}$ ,  $\beta_{MB}$ ,  $\omega_{train}$  and  $\omega_{gen}$ . For Experiment 3, we considered a four-weight hybrid model which has one weight for the Multiple trials, one weight for the Single trials, one weight for the first generalization phase and one weight for the second generalization phase, giving a total of seven free parameters.

*Arbitration model (AR):* Based on preliminary modeling results, where we fitted an unconstrained model using a time series of weights as free parameters (Figure S2.7), we considered that the mixing weight,  $\omega$ , could change over time. This observation aligns with previous research suggesting a transition from model-based to model-free control, which underlies the formation of habits (Gläscher et al., 2010). Accordingly, we propose an arbitration mechanism where the weight of the model-based component varies as a function of the history with familiar and novel states. In particular, we use a Bayesian updating approach, where the model-based weight on every time step  $\omega_t^{MB}$  is given by a Beta distribution:

$$\omega_t^{MB} \sim \text{Beta}(\alpha_t, \beta_t),$$

where

$$\alpha_t = \alpha_{t-1} + [S_{t-i} \neq S_t \text{ for } i \in \{1, 2, \dots, \tau\}]$$

$$\beta_t = \beta_{t-1} + [S_{t-i} = S_t \text{ for } i \in \{1, 2, \dots, \tau\}]$$

$S$  is the state, encoded as the current location and the target location. In this update rule, the parameter  $\alpha$  of the Beta distribution increases by one unit if the current state has not been experienced in the past, with the memory window parameter  $\tau$  controlling how far into the past to consider. If novel states are continuously being experienced, the distribution of  $\omega^{MB}$  will shift towards 1, increasing the control of the model-based system. On the other hand, if the current state has been experienced in the past, the parameter  $\beta$  of the Beta distribution increases by one unit. Therefore, If familiar states are continuously being experienced, the distribution of  $\omega^{MB}$  will shift toward 0, increasing the control of the model-free system. In order to mix the output of the model-based and the model-free system at every time step we use the mean of the Beta distribution  $\omega_t = \frac{\alpha_t}{\alpha_t + \beta_t}$ :

$$\phi_t^{Hybrid} = \omega_t \phi_t^{MB} + (1 - \omega_t) \phi_t^{MF}$$

$$R_t \sim Cat(\phi_{t-1}^{Hybrid [1]}, \phi_{t-1}^{Hybrid [2]}, \phi_{t-1}^{Hybrid [3]}).$$

This arbitration model provides a simple mechanism of how the weight of the model-based system could differ between the Single and Multiple group across the experiment (Figure S2.8). In particular, the frequent changes in the start-target pairs for the Multiple group makes it more likely that previously visited states are more distant in the past, potentially being subject to forgetting (as modulated by the  $\tau$  parameter). Therefore, resulting in greater model-based weights during training and at the beginning of the generalization phase. In contrast, for the Single group, familiar states are more likely to be found in the recent past due to the lack of changes in the start-target pair. This results in lower model-based weights during the training phase and at the beginning of the generalization phase.

This model has four free parameters:  $\alpha$ ,  $\beta_{MF}$ ,  $\beta_{MB}$  and  $\tau$ .

*Model fitting and evaluation:* We used Bayesian Adaptive Direct search (Acerbi, 2017) implemented in Matlab code to obtain point estimates of the parameters of our models. For each participant and model, we computed the Akaike information Criterion (AIC; Akaike, 1947) and the Bayesian Information Criterion (BIC; Schwarz, 1978).

*Absolute goodness of fit:* In addition to computing the AIC and BIC, which allows us to compare models among themselves, we wanted to see how well they described the data in the absolute sense, that is, compared to a theoretical (near) upper boundary for any probabilistic model, at least given particular assumptions about exchangeability. This approximate upper limit is represented by the negative entropy (Shen and Ma, 2016; Grassberger, 1988, 2003) and is given by:

$$-H(p(D|M_{true})) = \sum_D p(M_{true}) \log p(M_{true}),$$

where  $p(D|M_{true})$  represents the probability distribution of the data given the true model.

The negative entropy is a non-positive quantity and intuitively represents how much we can know about the data from the true generative model. An estimator of the negative entropy that has small error even with few data points is given by Grassberger (2003). For our experiments, this estimator is given by:

$$-\hat{H}(p(D|M_{true})) = \sum_{i=1}^C N_i (G_{N_i} - \frac{1}{N_i} (K_i^{[1]} G_{K_i^{[1]}} + K_i^{[2]} G_{K_i^{[2]}} + K_i^{[3]} G_{K_i^{[3]}})),$$

$$\text{with } K_i^{[1]} + K_i^{[2]} + K_i^{[3]} = N_i, \quad ,$$

where  $G_0 = 0$ ,  $G_1 = -\gamma - \log 2$ , and  $G_2 = 2 - \gamma - \log 2$ .  $\gamma \approx 0.577215$  is Euler's constant.

For ( $n \geq 1$ ):

$$G_{2n+1} = G_{2n},$$

$$G_{2n+2} = G_{2n} + \frac{2}{2n+1},$$

Thus,

$$n = 2$$

$$G_{2n} = - \underbrace{\gamma - \log 2 + \frac{2}{1}}_{n=1} + \frac{3}{2} + \dots + \frac{2}{2n+1}$$

$$n = 1$$

$C$  represents partitions of the data, e.g., experimental conditions, which in our case equals the number of unique pairs of start and target locations times the number of states in the grid. Therefore,  $C$  was not the same in all the experimental groups. In Experiment 1, the number of states in the grid was 81. The number of start and target pairs was five for the Single group and eight for the Multiple group. Therefore,  $C = 405$  and  $C = 648$  for the Single and Multiple groups, respectively. In Experiment 2 and Experiment 4, the number of grid states was 81 for both groups. The number of unique pairs of start and target locations was thirteen for the Single group and sixteen for the Multiple group. Therefore,  $C = 1053$  in the Single group and  $C = 1296$  in the Multiple group. Lastly, for Experiment 3 there were seventeen unique pairs of start and target locations across the experiment, therefore  $C = 1377$ .  $N_i$  is the total number of responses in the partition  $i$  of the data.  $K_i^{[1]}$  is the number of responses to key 1 in the  $i$  partition of the data,  $K_i^{[2]}$  the number of responses to key 2 in the  $i$  partition of the data and  $K_i^{[3]}$  the number of responses to key 3 in the  $i$  partition of the data. Importantly, this estimator assumes that the distribution of the data given the true model is stationary, which is not necessarily the case of our task as participants' responses can change due to learning. However, given that subjects'

performance stabilized relatively quickly as we can see in Figure 2.2 and 2.5, we considered it would be a reasonable approximation to an upper boundary of the models' performance.

The negative entropy can be compared with the negative cross-entropy, which intuitively represents how much we can know about the data from an imperfect model (our models). The negative cross entropy is given by:

$$-H(p(D|M_{true}), p(D|M_i)) = \sum_D p(D|M_{true}) \log p(M_i),$$

where  $p(D|M_i)$  represents the probability distribution of the data given the proposed model  $M_i$ .

The negative cross-entropy is also a non-positive value. An estimator of the negative cross entropy is the logarithm of the likelihood function evaluated at the maximum likelihood estimates of the parameters (Shen and Ma, 2016), which is returned by our parameter estimation method (Acerbi and Ma, 2017). In order to provide a simple visualization of the absolute goodness of fit, we computed the proportion of the explainable variability in the data that was explained by the models:

$$Proportion\ explained = \frac{H(p(D|M_{true}), p(D|M_i)) + \log p(D|M_{rand})}{H(p(D|M_{true})) + \log p(D|M_{rand})},$$

$\log p(D|M_{rand})$  is the logarithm of the likelihood of the data given a model that assumes all responses are equally likely and represents a lower boundary for all models. In the numerator, we have what *is* explained by a proposed model (as compared to the lower boundary), relative to what *can* be explained (difference between the upper and lower boundary), which is in the denominator.

**Group model comparison:** In addition to comparing the models among themselves at the individual level, we performed group model comparison following Stephan et al. (2009). Based on their work, the probabilities ( $q_1, q_2, q_3$ ) of our models in the population follow a Dirichlet distribution:

$$(q_1, q_2, q_3) \sim Dir([\alpha_1, \alpha_2, \alpha_3]),$$

the parameters  $\alpha = [\alpha_1, \alpha_2, \alpha_3]$  can be estimated by iterating the following algorithm provided by the authors and that we implemented in R code:

$$\alpha = [1, 1, 1]$$

*Until convergence:*

$$u_{nk} = \exp(\log p(y|M_{nk}) + \psi(\alpha_k) + \psi\left(\sum_k \alpha_k\right))$$

$$\beta_k = \frac{\sum u_{nk}}{n \sum_k u_{nk}}$$

$$\alpha = \alpha_0 + \beta$$

*end,*

where  $k$  is the number of tested models,  $n$  the number of subjects and  $\psi$  the digamma function. Importantly, this algorithm only requires that we provide the log marginal likelihood which can be approximated as  $-\text{BIC}/2$ . In order to avoid extremely big numbers from  $u_{nk}$ , which returned  $\infty$  in R, we used the logarithm of the marginal likelihood with base 100. We iterated the algorithm  $10^3$  times to provide reliable estimates of  $\alpha$ . The new parameters of the Dirichlet distribution can be used to compute the probabilities  $[r_1, r_2, r_3]$  that a randomly selected subject follows any of the tested models:

$$[r_1, r_2, r_3] = \left[ \frac{\alpha_1}{\tau}, \frac{\alpha_2}{\tau}, \frac{\alpha_3}{\tau} \right], \text{ with } \tau = \sum_{i=1}^3 \alpha_i$$

Finally, we computed the probability that a given model  $k$  is more likely than the others in the population, i.e., the exceedance probability  $\phi_k$ :

$$\forall j \in \{1, 2, 3 \mid j \neq k\},$$

$$\varphi_j = p(q_j > q_k | \alpha),$$

by the Law of Total Probability:

$$\varphi_j = \int_0^{\infty} p(q_j > q_k | q_j, \alpha) p(q_j | \alpha) dq_j$$

This integral can be approximated numerically by the method provided in Soch and Allefeld (2016) implemented in Matlab code.

*Parameter recovery:* In order to verify that the parameters from the tested models were identifiable, we generated data from each of the models using 100 random samples of their parameter space; then, we performed maximum likelihood estimation using Bayesian Adaptive Direct Search (Acerbi and Ma, 2017) to attempt to recover the parameters generating the data. Finally, we plotted the simulated versus the fitted parameters and computed the Pearson correlation between them. As can be observed in Figure S2.11, we were able to recover the simulated parameters reasonably well in all cases.

*Model recovery:* We simulated 100 data sets from each of the models using random samples of the parameter space. Then fitted those data sets with all the models. In Figure S2.12 we showed the confusion matrix with the proportion of times that each model was able to recover the data generated by the other models the best according to the Bayesian Information Criterion. In general, all the models were able to recover their own data better than the other models above 85% of the times.

## Results

### Experiment 1

How does training variability constrain learning and generalization of a visuomotor mapping? Two groups of participants performed a grid navigation task (Figure 2.1A) where they

moved a cursor from various start to target locations using the J, K and L keys of a standard keyboard (see *Materials and Methods* for details). In the Single group (n=16), participants trained to move between a single start-target pair, while in the Multiple group (n=16) participants trained to move between four start-target pairs (Figure 2.1B). We predicted that performance improvements for the Multiple group would be slower during training, but they would be able to generalize their performance to novel start-target pairs, reflecting the formation of a key-to-direction mapping rather than local state-action associations. In contrast, the Single group would show faster performance improvements during training, but be unable to generalize to new start-target pairs. Participants performed 260 training trials followed by 20 trials of generalization interleaved with 20 training trials.

### ***Behavioral Results***

Figure 2.2A shows the proportion of optimal arrivals over trial bins (1 bin = 10 trials) for both groups, where an optimal arrival is a trial where subjects arrive at the target using the minimum number of key presses (7 moves). As expected, the Multiple group had a slower learning curve as revealed by a mixed-effects model analysis. Specifically, the Single group demonstrated significantly higher performance compared to the Multiple group up to the bin corresponding to trial 60 ( $p < 0.05$ ). However, both the Single and Multiple groups reached the same level of performance by the end of the training phase (comparison of the last bin of training trials between groups;  $t(29.06) = -0.5$ ,  $p = 0.61$ ; Figure 2.2B). In addition, in the Multiple group, there were no substantial differences in performance among the different start-target pairs, with optimal arrivals generally increasing over time (Figure S2.1). Of more importance is how well the groups perform when new start-target pairs are introduced in the generalization phase. This was determined by quantifying any potential change in performance between the end of the training phase and beginning of the generalization phase. For the Single group, we found that performance was significantly worse at the onset of the generalization phase

compared to the end of the training phase ( $t(15) = -5.67$ ,  $p < 0.001$ ; Figure 2.2D). In contrast, the Multiple group's performance during generalization was similar to the end of the training phase ( $t(15) = -1.21$ ,  $p = 0.242$ ). When comparing between the groups, it was clear that the Multiple group performed significantly better than the Single group early in generalization (first bin;  $t(29.99) = -5.1$ ,  $p < 0.001$ ; Figure 2.2C) and still marginally better at the end of the generalization (second bin;  $t(29.27) = -1.64$ ,  $p = 0.05$ ). The change in performance between the groups was significantly different ( $t(29.34) = -3.5$ ,  $p = 0.001$ ; Figure 2.2D).

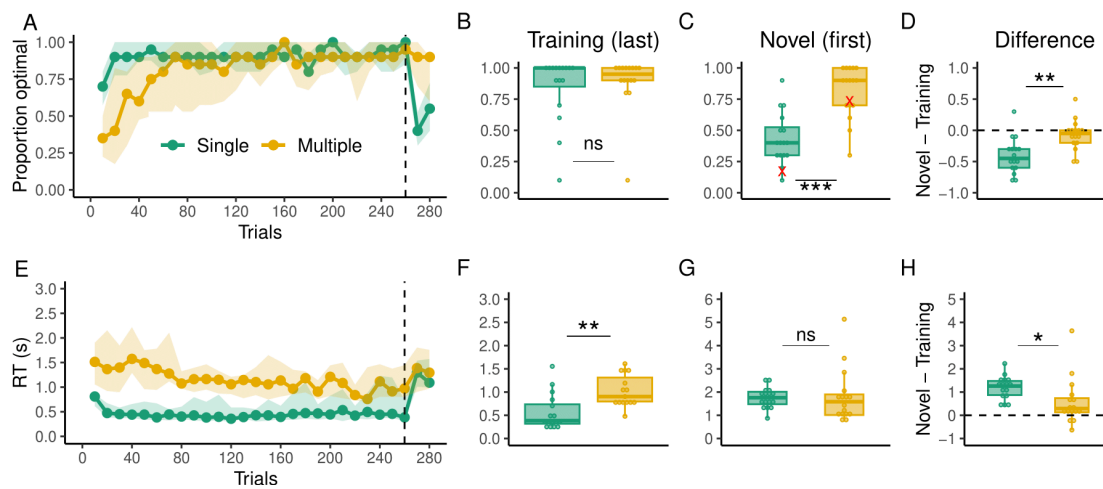


Figure 2.2: Behavioral results of Experiment 1. **(A)** Proportion of optimal arrivals over trial bins for the Single (green) and Multiple (gold) groups. The black dashed line indicates the beginning of the generalization phase. The solid dotted line represents the median and the shading region the interquartile range. **(B)** Proportion of optimal arrivals in the last bin of the training phase. **(C)** Proportion of optimal arrivals in the first bin of generalization (novel pairs). Red marks indicate performance in the very first trial of generalization for all subjects. **(D)** Difference in the proportion of optimal arrivals between the first bin of generalization and the last one of training. The dashed line here indicates no performance change from training to generalization **(E)** RTs over trial bins. **(F)** RTs in the last bin of training trials. **(G)** RTs in the first bin of generalization **(H)** Difference in RTs between the first bin of generalization and the last one of training.

We also examined how planning (reaction time, RT) evolved over training. RTs were overall higher in the Multiple group group throughout training ( $t(29.99) = -4.82$ ,  $p < 0.001$ ; Figure

2.2E) and higher at the last bin of training trials ( $t(21.21) = -2.85$ ,  $p = 0.009$ ; Figure 2.2F) but not during the first bin of generalization trials ( $t(19.36) = -0.17$ ,  $p = 0.86$ ; Figure 2.2G). There was an increase in RTs from the last training bin to the first generalization bin in the Single ( $t(15) = 9.54$ ,  $p < 0.001$ ; Figure 2.2H) and Multiple groups ( $t(15) = 2.34$ ,  $p = 0.03$ ), however, this increase was significantly greater in the Single group ( $t(21.96) = 2.14$ ,  $p = 0.043$ ). In addition, the groups only differ in RTs over training, but not in any of the inter-key-intervals (i.e., the time in between key presses; Figure S2.2). Crucially, however, we found that during the generalization phase, the Single group had significantly higher inter-key-intervals than the Multiple group (Figure S2.3), suggesting that they had to replan at intermediate steps of the trajectory unlike the Multiple group. For the Multiple group, we did not find substantial differences in terms of RTs among the different start-target pairs in terms of RTs. In general, RTs decreased over time in all pairs (Figure S2.1).

Differences in generalization performance provide preliminary evidence that the Multiple group could have used a more flexible representation of the mapping compared to the Single group to arrive at the novel targets. In the following analysis we asked whether a similar distinction could have arisen during the training phase. To investigate this, we hypothesized that participants who rely on the visuomotor mapping would make more frequent choices that require planning, compared to participants relying primarily on state-action values. In particular, the former would be more inclined to choose options where the cursor counterintuitively moves away from the target while actually reducing the true, model-based distance (Figure 2.3A). While choices that move the cursor visually closer to the target can result from heuristics that minimize unconstrained distance metrics (e.g., the chessboard distance), choices that minimize only the model-based distance (“MB move”) require considering the long-term outcomes of using the constrained mapping.

Indeed, we found that participants in the Multiple group made more MB moves in the training phase compared to the Single group ( $t(29) = -4.34$ ,  $p < 0.001$ ; Figure 2.3A left), and also

relative to the total number of moves ( $t(28.17) = -5.26$ ,  $p < 0.001$ ; Figure 2.3A middle), which controls that participants in the Multiple group made on average more key presses in the experiment. Most notably, we showed that this pattern remained when considering the opportunities that participants had to minimize the model-based distance over the chessboard distance ( $t(23.46) = -2.34$ ,  $p = 0.02$ ; Figure 2.3A right). The latter analysis excludes situations where minimizing the chessboard distance or the model-based distance are confounded. For example, when the target is only one move away from the cursor, the winning move minimizes both distances while the other two moves increase them; similarly, when the target is to the right of the cursor and on the same row, pressing the key that moves the cursor to the right is the only move that minimizes both distances while the other two moves increase them. Such situations do not represent a choice opportunity in our analysis. When participants in the Multiple group had the opportunity to minimize either the model-based distance or the chessboard distance, they chose to minimize the former more often than chance (i.e., more than 50% of the times;  $t(15) = 3.09$ ,  $p = 0.007$ ), while in the Single group they did not ( $t(15) = -0.95$ ,  $p = 0.35$ ).

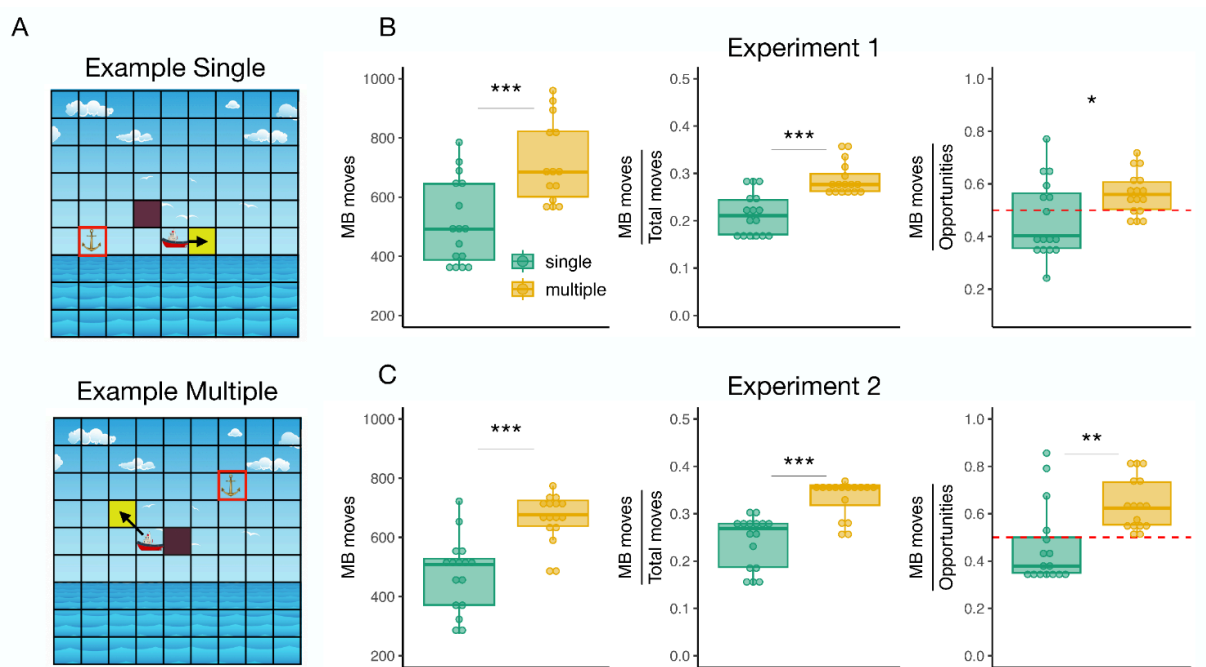


Figure 2.3: Behavioral evidence of model-based computations during training of Experiment 1 and Experiment 2. (A) Examples in the Single (top) and Multiple (bottom) groups where the cursor can move to two states that minimize the distance to the target. While moving to the purple state minimizes both the chessboard distance (getting visually closer to the target) and the model-based distance (getting closer to the target in mapping space), moving to the yellow state (MB move) minimizes only the true, model-based distance, i.e., the cursor moves visually away from the target but closer in mapping space. (B) Left: Total number of MB moves during training for each group. Middle: Number of MB moves relative the total number of moves in the experiment. Right: Proportion of MB moves relative to the number of opportunities to make that choice, i.e., choosing to minimize the model-based distance over the chessboard distance as in (A). (C) The same as in (B) but for Experiment 2. The red dashed lines indicate chance level when choosing to minimize the chessboard distance or the model-based distance.

In addition, we also examined the most frequently used trajectories across participants during the training phase (Figure S2.4). We found that the preferred trajectories were less likely to be shared by other participants in the Single group compared to the Multiple group. In particular, between 3-5 participants (18-31%) in the Single group shared the same preferred trajectory. In contrast, between 7-11 participants (43-68%) in the Multiple group converged to the same solution for the majority of the trials. The greater heterogeneity in the trajectories used by the Single group likely resulted from the requirement to utilize all three keys to arrive at the target. On the other hand, the Multiple group only required the use of two keys, although all keys were needed across the experiment. In fact, using a dynamic programming approach, we determined that there were a total of 140 trajectories that optimally reached the target in the Single group, whereas in the Multiple group, there were only between 21 and 35 optimal trajectories depending on the start-target pair. The larger number of successful trajectories in the Single group likely contributed to fewer participants finding the same solutions.

In a related analysis, we found that, within participants, both the Single and Multiple groups showed highly homogeneous trajectories during the training phase. In particular, participants in the Single group reliably chose two trajectories more than 50% of the time ( $t(15) = 2.86$ ,  $p = 0.005$ ; Figure S2.5), whereas the Multiple group exclusively selected a single

trajectory more than 50% of the time for each of the four start-target pairs ( $t(15) = [3.28, 2.56, 2.54, 2.05]$ ,  $p = [0.002, 0.01, 0.01, 0.02]$ ). While both groups displayed highly repetitive behavior relative to their target, the absolute frequency of repetitions of the preferred trajectories was much smaller in the Multiple group given that, by design, each target appeared for fewer trials. Indeed, the most frequently used trajectory in the Single group was used significantly more times than the maximum number of times the trajectory in the Multiple group could have been used (i.e., 25% of the time;  $t(15) = 4.36$ ,  $p < 0.001$ ).

The differences in generalization performance between the Single and Multiple groups, which remained marginally significant even by the end of the generalization phase, suggest that the training regimes might have led to different learning representations in the task. While both groups chose highly repetitive trajectories during training, which aligns with the memorization of few solutions, the variability experienced by the Multiple group could have simultaneously allowed the better learning of the visuomotor mapping. Evidence supporting this comes from the more frequent selection of moves that exclusively reduce the model-based distance by the Multiple group. In addition, the greater repetition of the preferred trajectories in the Single group given its unique target, could have facilitated the formation of habitual responses which interfered in the generalization phase. Indeed, supplementary analysis revealed that errors in the generalization phase of the Single group more often began with the same key as the one most frequently used trajectory during the training phase, compared to the Multiple group (Figure S2.6). While these differences are numerical due to the limited power resulting from the low error rate in the Multiple group, they suggest that beside a greater representation of the mapping, training variability could have also prevented the interference from habitual responses.

### ***Modeling Results***

In order to explore the cognitive processes that gave rise to the differences in performance between the groups, we evaluated five computational models from the

reinforcement learning literature (see *Materials and Methods* for details). On one end of the spectrum, we tested a model-free reinforcement learning algorithm (Sutton and Barto, 1998) which learns state-action values using prediction errors based on the chessboard distance to the target. On the other end of the spectrum, we implemented a model-based algorithm that learns the visuomotor mapping of the task and uses it to find the shortest route to the target using tree search. We selected these models as they make contrasting predictions about generalization performance in the task (Figure 2.4). A fully trained model-free reinforcement learning algorithm would have good performance for familiar target locations, but generate chance-level responses for targets that have not been experienced in the past, thus predicting poor generalization. A model-based algorithm would similarly perform well for familiar targets but would instead be able to generalize to novel targets once the visuomotor mapping has been learned. Finally, we tested hybrid models between the model-free and model-based algorithms, which are weighted sums of the two components. Hybrid models would perform well in familiar targets as the model-free and model-based algorithms, but would be able to capture different degrees of generalization (Figure 2.4).

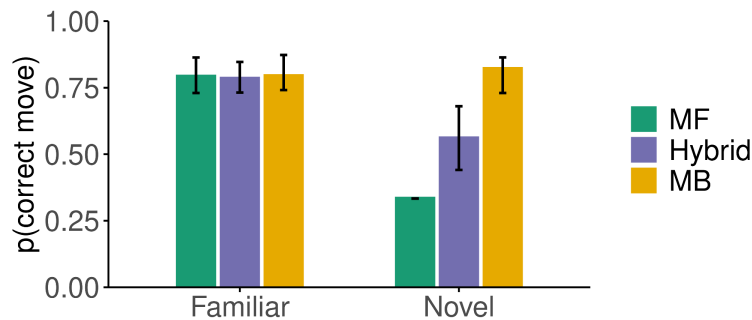


Figure 2.4: Models simulation. Probability of choosing the optimal move for the fully trained models on a familiar and novel target. All models are capable of achieving ceiling performance for familiar targets after sufficient training similar to human subjects. Responses were generated using an inverse temperature parameter  $\beta = 4$  for all models. For the hybrid model 100 Uniform(0,1) samples of the model-based weight were used to simulate its performance. The height of the bars represents the median and the error bars the interquartile range.

Based on previous work (Daw et al., 2011; Gläscher et al., 2010; Kool, 2017; Lee et al., 2014), we considered that the mixing weight between the model-free and model-based components in the hybrid model could take the following forms: a single weight across the entire experiment (1W), one weight for the training phase and one for the generalization phase (2W), and a time varying weight. For the latter model, we developed a Bayesian arbitration mechanism based on the history of familiar or novel states (AR model; see *Materials and Methods* for details) which assigns a greater weight to the model-based component if novel states are frequently encountered. On the other hand, if familiar scenarios are experienced, the model-based weight decreases, giving way to the more habitual, model-free system. This transition from model-based to model-free control has been reported previously (Gläscher et al., 2010), and a preliminary modeling analysis revealed that this was a plausible scenario in our experiment given the task statistics of the Single and Multiple groups (Figure S2.7 and Figure S2.8). In order to compare the models, we used the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The results of the model comparison are shown in Table 2.1.

For the Single group, the 2W hybrid model best described the majority of participants according to both the AIC (14 out of 16) and the BIC (10 out of 16) metrics. In contrast, the Multiple group showed mixed results depending on the metric used. While the 2W hybrid model best described the majority of participants according to the AIC (11 out of 16), the BIC indicated that the arbitration hybrid model was the best for 7 out of 16 participants, followed by the 1W hybrid model (5 out of 16) and the 2W hybrid model (4 out of 16). To provide a global metric of model performance across subjects, we performed a model comparison at the group level (Stephan et al., 2009) see *Materials and Methods* for details; Figure 2.5A). This analysis indicated that the 2W hybrid model provided the best overall description for both groups. Specifically, a subject taken at random from the Single group had a 98% probability of being best described by the 2W hybrid model, followed by 2% from the arbitration model. For the

Multiple group, this probability was 57% for the 2W hybrid model, followed by 31% for the arbitration hybrid model, and 12% for the 1W hybrid model.

| Groups   | Models | AIC               | AIC wins | BIC               | BIC wins |
|----------|--------|-------------------|----------|-------------------|----------|
| Single   | MF     | 2637 [2037, 3563] | 0/16     | 2649 [2048, 3922] | 0/16     |
|          | MB     | 3968 [3339, 4555] | 0/16     | 3973 [3345, 4056] | 0/16     |
|          | 1W     | 2589 [1980, 3382] | 1/16     | 2612 [2003, 3665] | 1/16     |
|          | 2W     | 2556 [1968, 3359] | 14/16    | 2584 [1996, 3668] | 10/16    |
|          | AR     | 2585 [1965, 3381] | 1/16     | 2608 [1988, 3677] | 5/16     |
| Multiple | MF     | 3079 [2508, 3910] | 0/16     | 3090 [2519, 3922] | 0/16     |
|          | MB     | 3109 [2631, 4050] | 0/16     | 3115 [2637, 4056] | 0/16     |
|          | 1W     | 2766 [2272, 3641] | 1/16     | 2790 [2294, 3665] | 5/16     |
|          | 2W     | 2753 [2263, 3639] | 11/16    | 2782 [2291, 3668] | 4/16     |
|          | AR     | 2761 [2273, 3653] | 4/16     | 2784 [2296, 3677] | 7/16     |

Table 2.1: Individual model comparison for Experiment 1. The AIC and BIC columns show the median across subjects and the interquartile range inside the square brackets. AIC wins and BIC wins columns indicate the number of participants where the given model was the best one according to each metric.

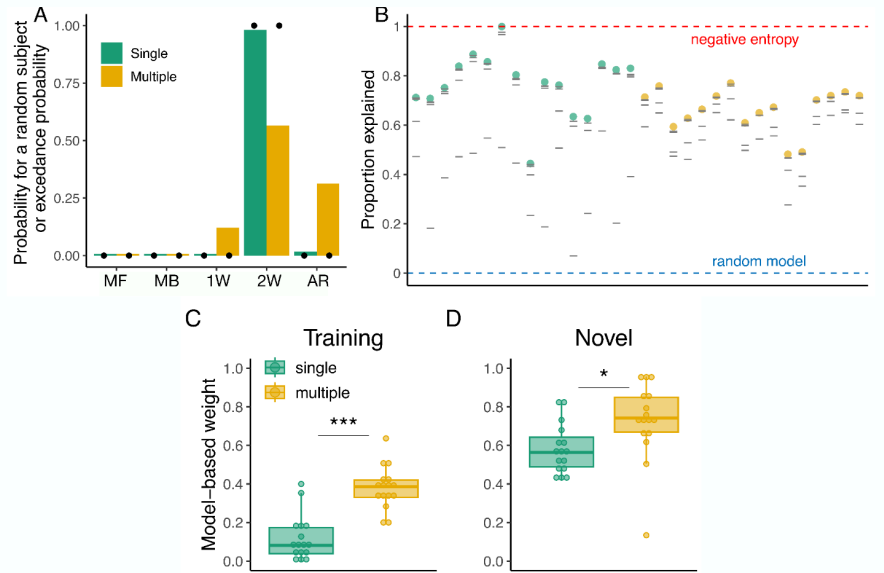


Figure 2.5: Modeling results of Experiment 1. **(A)**: Probability that a random subject taken from the Single and Multiple groups is best described by the tested models (MF: model-free, MB: model-based, 1W: Single-seight hybrid, 2W: Two-weight hybrid, AR: arbitration model). Black dots indicate the exceedance probability. **(B)** Proportion of the variability in the data explained by the models. Colored dots represent this value for the best model according to the BIC and gray lines represent the other models. Red and blue dashed lines represent the negative entropy (upper boundary) and the performance of a random model (lower boundary), respectively. **(C)** Model-based weight in the hybrid model during training and **(D)** generalization.

In addition, we computed the probability that a given model was more likely than the others in the population, i.e. the exceedance probability (Stephan, 2009). According to this metric, there was a high probability ( $>99\%$ ) that the 2W hybrid model was better than the rest of the models in both groups. In addition, in order to evaluate how good the models were in the absolute sense, we computed the proportion of the variability in our data that was explained by our models as compared to the negative entropy, which is a near upper bound for our probabilistic models (Shen and Ma, 2016; Grassberger, 1988, 2003; see *Materials and Methods* for details). The median proportion of the variability explained by the best model according to the BIC in the Single group was 76%, whereas in the Multiple group it was 68% (Figure 2.5B).

Given that the 2W hybrid model provided the best description of the data in the aggregate, we show the weights toward the model-based component during training and generalization for both the Single and Multiple groups (Figure 2.5C and 5D). Differences in the weights can indicate whether participants relied more on the model-free system during training due to repetitions but switched to the model-based system when the task demands changed. We performed a 2-Way ANOVA with the experimental phase (training or generalization) and the group (Single or Multiple) as factors, and found a significant main effect of the experimental phase ( $F(1,60) = 124.56$ ,  $p < 0.001$ ,  $\eta^2 = 0.56$ ), indicating that overall participants during the generalization phase had higher weights for the model-based component compared to the training phase. Similarly, we found a significant main effect of the group ( $F(1,60) = 31.9$ ,  $p < 0.001$ ,  $\eta^2 = 0.43$ ), where the Multiple group had higher model-based weights overall. A *post hoc* pairwise analysis using the Tukey's Honest Significant Difference (HSD) found that the Multiple group had significantly higher model-based weights during training (adjusted  $p < 0.001$ ) and generalization (adjusted  $p = 0.02$ ) compared to the Single group. These results corroborate our behavioral findings, where participants in the Multiple group tended to reduce the model-based distance to the target more often than the Single group during training (Figure 2.3B), and also that they had better generalization performance (Figure 2.2D).

Overall, the behavioral and modeling results from Experiment 1 suggest that greater variability during training enhances the use of the visuomotor mapping in novel scenarios, and potentially reduces the interference from habitual responses. These behavioral findings are supported by our computational modeling results, where we found greater model-based weights in the Multiple group. Nevertheless, it is relevant to note that other factors could have contributed to such results. In particular, the Multiple group could have shown better generalization performance if the trajectories used during training were more similar to the ones required to arrive at the targets during generalization as compared to the Single group. Second,

while the start-target pairs in generalization trials were novel, providing feedback about the cursor movement in such trials could generate learning about the mapping, which creates a confound between new learning and generalization. Lastly, it is unclear whether differences in generalization performance between the groups occurred because of a distinct representation at the level of the visuomotor mapping or because the groups differed in their ability to use the mapping sequentially (planning), with the Single group being unable to generate appropriate sequences with an otherwise known mapping. We address these concerns in Experiment 2.

## **Experiment 2**

In Experiment 1, we found that the Multiple group readily generalized their performance to new start-target pairs during the generalization phase, while the Single group struggled to generalize to novel pairs. However, feedback was available during the generalization phase, allowing the Single group to potentially relearn the mapping to recover performance, which is suggested by the elevated RTs during this phase. Here, we sought to control for this possibility by not providing feedback during the generalization phase and placing the target only one move away from the starting location (Figure 2.1B). In addition, to prevent further learning during the generalization phase, we also removed the interleaved training trials.

Removing the sequential component from the generalization phase can also rule out the potential confound that differences in performance arose due to distinct capabilities to plan new movement sequences resulting from the training regime. Finally, placing the target one step away, removes the possibility that better generalization performance occurs due to greater similarity between the sequences generated during training and the ones required in generalization.

If participants in the Single group still underperform the Multiple group in this simple situation, it would provide further evidence that they did not know or could not use the mapping to the extent that the Multiple group could by the end of the training phase. Apart from these

changes, everything else remained as in Experiment 1. There were a total of 260 trials of training for both the Single (n=16) and Multiple (n=16) groups and 20 trials of generalization.

### ***Behavioral Results***

Similar to Experiment 1, we found that the Multiple group had a slower learning curve (Figure 2.6A) as revealed by a mixed-effects model where the Single group had significantly higher performance up until the bin corresponding to trial 50 ( $p < 0.05$ ). However, both groups reached the same level of performance by the end of the training phase ( $t(29.99) = -0.19$ ,  $p = 0.84$ ; Figure 2.6B). In addition, no apparent difference in performance was noticed among the different start-target pairs of the Multiple group, with optimal arrivals generally improving with time (Figure S2.9). Of primary interest is how each group performed during the generalization phase where only one move was required and no feedback was provided. Notably, the Multiple group still outperformed the Single group early (first bin;  $t(22.68) = -4.81$ ,  $p < 0.001$ ; Figure 2.6C) but also late in the generalization phase (second bin;  $t(19.41) = -3.1$ ,  $p = 0.005$ ). However, the performance of the Single group remained greater than chance ( $t(15) = 5.88$ ,  $p < 0.001$ ), which suggests that even though they performed worse than the Multiple group, they could have recalled some knowledge about the mapping. In addition, the Single group's performance was significantly worse at the onset of generalization (first bin;  $t(15) = -5.03$ ,  $p < 0.001$ ; Figure 2.6D) compared to late in training, but the Multiple group's performance did not significantly decrease between the training and generalization phases ( $t(15) = -1.24$ ,  $p = 0.23$ ). The continued worse performance by the Single group in the generalization phase suggests that they struggled to effectively switch to model-based control and that variability in training may be necessary for a model-based algorithm to be efficiently used.

Similar to Experiment 1, RTs in the Multiple group were overall higher during training (Figure 2.6E;  $t(16.06) = -8.31$ ,  $p < 0.001$ ) and higher at the last bin of training trials (Figure 2.6F;  $t(28.13) = -4.22$ ,  $p < 0.001$ ) but not in the first bin of generalization trials (Figure 2.6G;  $t(26.48) =$

-1.08,  $p = 0.28$ ). RTs significantly increased from the last training bin to the first generalization bin in the Single ( $t(15) = 6.55$ ,  $p < 0.001$ ) but not in the Multiple group ( $t(15) = 0.41$ ,  $p = 0.68$ ). In addition, this change in RTs was significantly greater in the Single group ( $t(26.43) = 3.34$ ,  $p = 0.002$ ; Figure 2.6H). The differences in RTs increase suggest that the Multiple group did not experience a switch in the algorithm used to solve the task, whereas the Single group likely transitioned to a more computationally demanding algorithm. As in Experiment 1, we did not find noticeable differences in RTs among the different start-target pairs of the Multiple group, with RTs generally decreasing over time (Figure S2.9).

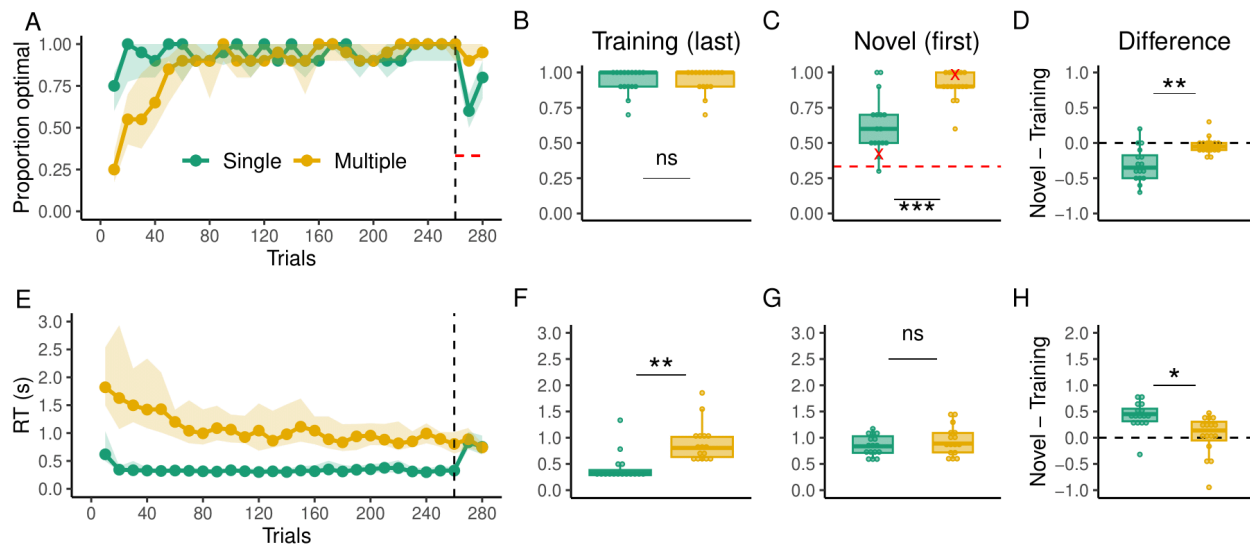


Figure 2.6: Behavioral results of Experiment 2. **(A)** Proportion of optimal arrivals over trial bins for the Single (green) and Multiple (gold) groups. The black dashed line indicates the beginning of the generalization phase. The solid dotted line represents the median and the shading region the interquartile range. The red dashed line indicates chance level of performance. **(B)** Proportion of optimal arrivals in the last bin of training trials. **(C)** Proportion of optimal arrivals in the first bin of generalization (novel pairs). Red marks indicate performance in the very first trial of generalization for all subjects. **(D)** Difference in the proportion of optimal arrivals between the first bin of generalization and the last one of training. The dashed line here indicates no performance change from training to generalization **(E)** RTs over trial bins. **(F)** RTs in the last bin of training trials. **(G)** RTs in the first bin of generalization **(H)** Difference in RTs between the first bin of generalization and the last one of training.

Subsequently, we performed the same analysis as in Experiment 1 looking for behavioral signals of the use of the visuomotor mapping over state-action associations during training. Corroborating our previous results, we found that participants in the Multiple group generated more moves that minimized exclusively the model-based distance to the target in the absolute sense ( $t(26.29) = -5.02$ ,  $p < 0.001$ ; Figure 2.3C left), relative to the total number of moves in the experiment ( $t(27.71) = -5.45$ ,  $p < 0.001$ ; Figure 2.3C middle) and relative to the opportunities they had to make this choice ( $t(25.66) = -3.55$ ,  $p = 0.001$ ; Figure 2.3C right). Likewise, we found that in situations where participants could minimize either the model-based distance or the chessboard distance, they chose to minimize the former significantly more than chance (50% of the times) in the Multiple group ( $t(15) = 5.12$ ,  $p < 0.001$ ; Figure 2.3C right) but not the Single group ( $t(15) = -0.92$ ,  $p = 0.37$ ).

In subsequent analyses, aiming to compare the generalization performance of participants in Experiment 1 (targets seven moves away) and Experiment 2 (targets one move away), we found converging evidence that participants in the Multiple group performed overall better. Specifically, we carried out a 2-Way ANOVA over the proportion of optimal arrivals across all generalization trials, with experiment number (Experiment 1 and Experiment 2) and group (Multiple and Single) as factors, which revealed a significant main effect of group ( $F(1,60) = 31.07$ ,  $p < 0.001$ ,  $\eta^2 = 0.29$ ; Figure 2.7A). Similarly, we found a significant main effect of experiment number ( $F(1,60) = 13.28$ ,  $p < 0.001$ ,  $\eta^2 = 0.12$ ), where participants in Experiment 2 performed overall better. *Post hoc* pairwise comparisons revealed that while there were significant differences in generalization between Experiment 1 and Experiment 2 in the Single group (adjusted  $p < 0.001$ ), this difference did not reach significant levels for the Multiple group (adjusted  $p = 0.09$ ), suggesting that planning primarily affected the optimal arrivals in the Single group.

We performed the same analysis over the RTs of Experiment 1 and Experiment 2, and found a significant main effect of experiment number ( $F(1,60) = 27.61$ ,  $p < 0.001$ ,  $\eta^2 = 0.31$ ; Figure 2.7B) but not of group ( $F(1,60) = 0.15$ ,  $p = 0.69$ ,  $\eta^2 = 0.04$ ). The *post hoc* pairwise comparisons revealed that participants in both the Single (adjusted  $p = 0.004$ ) and Multiple (adjusted  $p = 0.001$ ) groups had lower RTs in the generalization trials of Experiment 2, indicative of a lower computational demand compared to Experiment 1 where more planning was necessary.

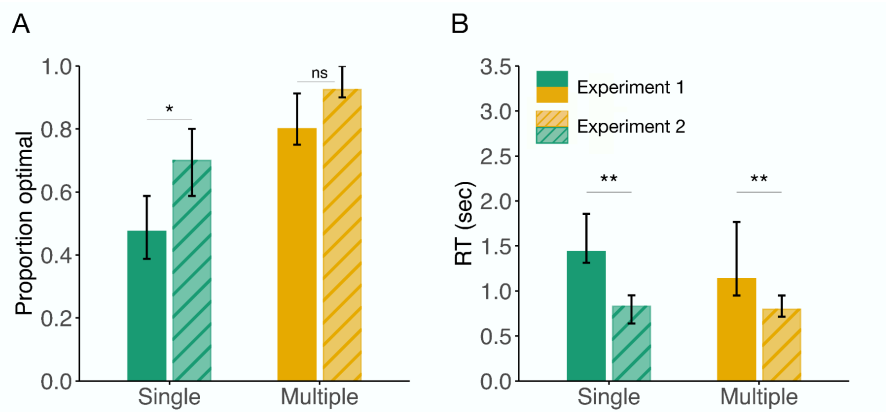


Figure 2.7: Comparison of generalization performance between Experiment 1 (smooth bars; target seven moves away) and Experiment 2 (striped bars; target one step away) for the Single and Multiple groups. **(A)** Proportion of optimal arrivals and **(B)** reaction times for all trials in the generalization phase. The height of the bars represents the median and the error bars the interquartile range.

Regarding the trajectories that participants used during the training phase, we found similar results as in Experiment 1. In particular, participants in the Single group converged to different trajectories with only 2-3 participants (12-18%) sharing their most frequently used trajectory. In contrast, 7-8 participants (44-50%) in the Multiple group shared the most frequent trajectory (Figure S2.4). As in Experiment 1, we observed repetitive behavior in both the Single and Multiple groups. Particularly, participants in the Single group chose a single trajectory more than 50% of the time ( $t(15) = 2.06$ ,  $p = 0.02$ ; Figure S2.5). Similarly, the Multiple group selected

a single trajectory more than 50% of the time ( $t(15) = [6.21, 6.23, 5.63, 3.91]$ ,  $p < 0.001$ ). However, as in Experiment 1, the absolute frequency of repetitions for preferred trajectories was much smaller in the Multiple group given that each target appeared for fewer trials. The most frequently used trajectory in the Single group was used significantly more times than the maximum number of times the trajectory in the Multiple group could have been used (i.e., 25% of the time;  $t(15) = 5.88$ ,  $p < 0.001$ ).

### **Modeling Results**

As in Experiment 1, we evaluated five models: the model-free, model-based and three hybrid models. The results of the model comparison are presented in Table 2.2.

| Groups   | Models | AIC               | AIC wins | BIC               | BIC wins |
|----------|--------|-------------------|----------|-------------------|----------|
| Single   | MF     | 2090 [1526, 2603] | 1/16     | 2101 [1537, 2614] | 3/16     |
|          | MB     | 3204 [2775, 3526] | 0/16     | 3209 [2780, 3532] | 0/16     |
|          | 1W     | 2074 [1498, 2482] | 1/16     | 2096 [1520, 2505] | 2/16     |
|          | 2W     | 2069 [1470, 2465] | 9/16     | 2097 [1497, 2493] | 5/16     |
|          | AR     | 2072 [1484, 2479] | 5/16     | 2094 [1507, 2502] | 6/16     |
| Multiple | MF     | 2468 [2063, 2854] | 0/16     | 2497 [2074, 2865] | 0/16     |
|          | MB     | 2436 [2164, 2780] | 0/16     | 2441 [2169, 2785] | 0/16     |
|          | 1W     | 2157 [1845, 2551] | 1/16     | 2197 [1868, 2573] | 3/16     |
|          | 2W     | 2132 [1838, 2543] | 14/16    | 2160 [1865, 2571] | 8/16     |
|          | AR     | 2156 [1860, 2580] | 1/16     | 2178 [1882, 2602] | 5/16     |

Table 2.2: Individual model comparison for Experiment 2. The AIC and BIC columns show the median across subjects and the interquartile range inside the square brackets. AIC wins and BIC wins columns indicate the number of participants where the given model was the best one according to each metric.

According to the AIC, we found that the majority of participants were best described by the 2W hybrid model in the Single (9 out of 16) and Multiple groups (14 out of 16). However, some participants were best described by the rest of the models as well. In contrast, according to the BIC metric, there was less evidence for a dominant model at the individual level. In the Single group, the arbitration model best described the largest number of participants (6 out of 16), followed by the 2W hybrid model (5 out of 16), the model-free model (3 out of 16), and the 1W hybrid model (2 out of 16). For the Multiple group, the 2W hybrid model best described half of the participants (8 out of 16), followed by the 1W hybrid model (5 out of 16) and the arbitration model (3 out of 16).

When we considered the aggregate evidence for each model across participants, we found that a random subject taken from the Single group would have a 66% probability of being best described by the 2W hybrid, 27% by the arbitration model, 5% by the 1W hybrid model and 2% by the model-free model. For the Multiple group, there was a 92% probability that a random subject would be best described by the 2W hybrid model, followed by 6% from the 1W hybrid model and 2% from the arbitration model. According to the exceedance probability, there was a high probability (>99%) that the 2W hybrid model was better than the rest of the models in both the Single and Multiple groups. Thus, at the population level, the 2W hybrid was the predominant model overall (Figure 2.8A). For the best performing models at the individual level according to the BIC, we found that, when compared with the theoretical near upper bound, they captured 73% and 68% of the explainable variability in the data in the Single group and Multiple group, respectively (Figure 2.8B).

Similar to Experiment 1, we analyzed the weights of the 2W hybrid model during training and generalization. A 2-Way ANOVA with experimental phase (training or generalization) and group (Single or Multiple), showed a significant main effect of the experimental phase ( $F(1,60) = 111.8$ ,  $p < 0.001$ ,  $\eta^2 = 0.48$ ) and the group ( $F(1,60) = 60$ ,  $p < 0.001$ ,  $\eta^2 = 0.26$ ), indicating that

overall participants had higher model-based weights during generalization and in the Multiple group. Post hoc pairwise comparisons revealed that the model-based weight was significantly higher in the Multiple group during training (adjusted  $p < 0.001$ ; Figure 2.8C) and generalization (adjusted  $p < 0.001$ ; Figure 2.8D). These results provide further evidence that the observed behavioral differences between the groups during training (Figure 2.3C) and generalization (Figure 2.6C) can be described by the components of our model.

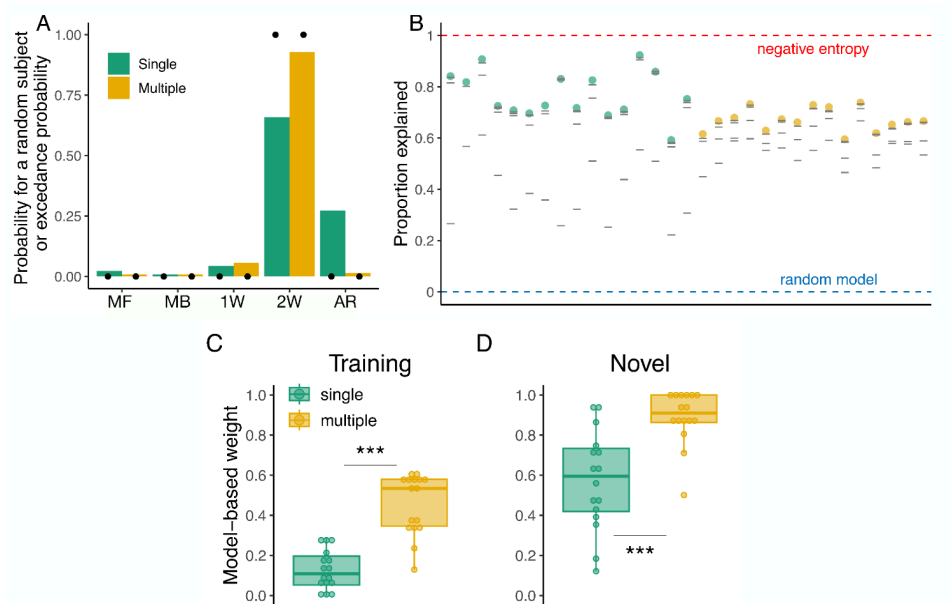


Figure 2.8: Modeling results of Experiment 2. **(A)** Probability that a random subject taken from the Single and Multiple groups is best described by the tested models (MF: model-free, MB: model-based, 1W: Single-seight hybrid, 2W: Two-weight hybrid, AR: Arbitration model). Black dots indicate the exceedance probability. **(B)** Proportion of the variability in the data explained by the models. Colored dots represent this value for the best model according to the BIC and gray lines represent the other models. Red and blue dashed lines represent the negative entropy (upper boundary) and the performance of a random model (lower boundary), respectively. **(C)** Model-based weight in the hybrid model during training and **(D)** generalization.

When taken together, the behavioral and modeling results from Experiments 1 and 2 indicate that training variability encouraged a robust learning of the visuomotor mapping, which manifested in successful generalization. While the Single group still shows evidence of having

learned the mapping, they struggle to generalize even in simple situations that do not require sequential planning. In contrast, the Multiple group maintained ceiling performance regardless of whether planning was necessary (Figure 2.7). In contrast to Experiment 2, differences in generalization held true when withholding feedback and removing the sequential component in the generalization trials – therefore ruling out the confound of relearning and planning. In addition, those differences remained significant until the end of the generalization phase.

Further supplementary analysis also revealed numerical differences indicating that errors during generalization were more likely to be the result of interference from habitual responses in the Single group than in the Multiple group (Figure S2.6). Our behavioral results were supported by our modeling analysis, where a hybrid model with separate weights for the training and generalization phases provided the best description of the data, and significant differences in its weights mirrored the behavioral differences during training and generalization between the Single and Multiple groups (Figure 2.3C and Figure 2.6C).

In the following experiment, we inquire whether the benefits in generalization performance due to training variability can be retained over time, even when followed by a long exposure of repetitive practice with no variability. If this were the case, this would provide evidence that early formation of the visuomotor mapping can reduce the interference of habitual responses in the future and allow for flexible behavior.

### **Experiment 3**

In the previous experiments, we demonstrated that the Single group's performance consistently decreased during the generalization phase. In this study, we investigated whether a short exposure to training variability could prevent this decline, even when followed by a prolonged period with no variability. To test this idea, 16 participants were briefly trained with four start-target pairs (80 trials; now Multiple trials), followed by a first generalization phase (20 trials; Figure 2.1B). Subsequently, they were exposed to a single start-target pair for a prolonged

period of time (1000 trials; now Single trials), followed by a second generalization phase (20 trials). As in Experiment 2, the target locations were placed one move away from the starting point and no feedback was provided. If a visuomotor mapping is trained early on and maintained in memory – even if not being used – then participants would show good performance both in the first and second generalization phase. Alternatively, the prolonged period with no variability in the Single trials, most likely dominated by model-free processes, could impair the previously learned mapping resulting in good performance in the first but not in the second generalization phase.

### ***Behavioral results***

We found that prior to starting each generalization phase, the performance of participants was not statistically different between the Single trials and Multiple trials ( $t(15) = 1.58$ ,  $p = 0.13$ ; Figure 2.9B). Most notably, generalization performance was not significantly different after the Single and Multiple trials ( $t(15) = 1.05$ ,  $p = 0.3$ ; Figure 2.9C). Similarly, there was no significant change in performance from the end of the Multiple trials to the beginning of the first generalization phase ( $t(15) = 1.05$ ,  $p = 0.3$ ), nor from the end of the Single trials to the beginning of second generalization phase ( $t(15) = -1.43$ ,  $p = 0.17$ ). Furthermore, we did not find a significant difference among the change in performance in the two trial phases ( $t(29.52) = -1.88$ ,  $p = 0.07$ ; Figure 2.9D).

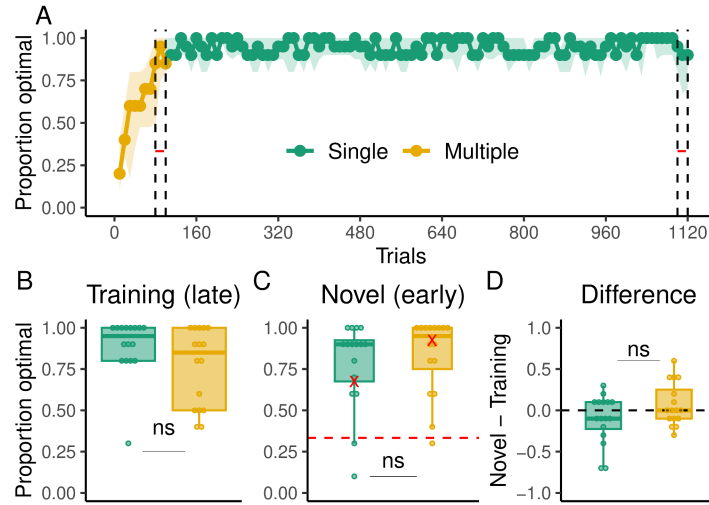


Figure 2.9: Behavioral performance in Experiment 3. **(A)** Proportion of optimal arrivals over trial bins. Participants were exposed to the Multiple trials (gold) followed by the Single group (green). The solid dotted line represents the median and the shading the interquartile range. The first (trials 81-100) and second (trial 1100-1120) generalization phases are demarcated with the dashed vertical lines. The red dashed line indicates chance level of performance. **(B)** Proportion of optimal arrivals in the last bin of training trials. **(C)** Proportion of optimal arrivals in the first bin of generalization (novel pairs). Red marks indicate performance in the very first trial of generalization for all subjects. **(D)** Difference in the proportion of optimal arrivals between the first bin of generalization and the last one of training.

In contrast with the comparable performance of participants in terms of optimal arrivals, their RTs differed considerably among the different stages of the experiment. In particular, RTs were significantly higher at the end of the Multiple trials compared to the end of the Single trials ( $t(15) = -3.28$ ,  $p = 0.003$ ; Figure 2.10B), suggesting that, as in our previous experiments, during the Multiple trials participants were relying more heavily on knowledge about the mapping (i.e., model-based computations). Crucially, we still found an increase in RTs from the end of Single trials to the beginning of the second generalization phase ( $t(15) = 5.18$ ,  $p < 0.001$ ; Figure 2.10D), suggesting a switch from state-action associations to model-based computations. However, in contrast to Experiment 1 and Experiment 2, such increase in RTs was accompanied by good generalization performance.

As in our previous experiments, we did not observe a change in RTs from the end of the Multiple trials to the beginning of the first generalization phase ( $t(15) = -1.52$ ,  $p = 0.14$ ; Figure 2.10D), which suggest that participants did not have to switch to a different algorithm during this transition. Notably, given the extended practice in the Single trials, all the inter-key-intervals were significantly lower ( $p < 0.01$ ) than in the Multiple trials (Figure S2.2), which was not observed in the previous experiments.

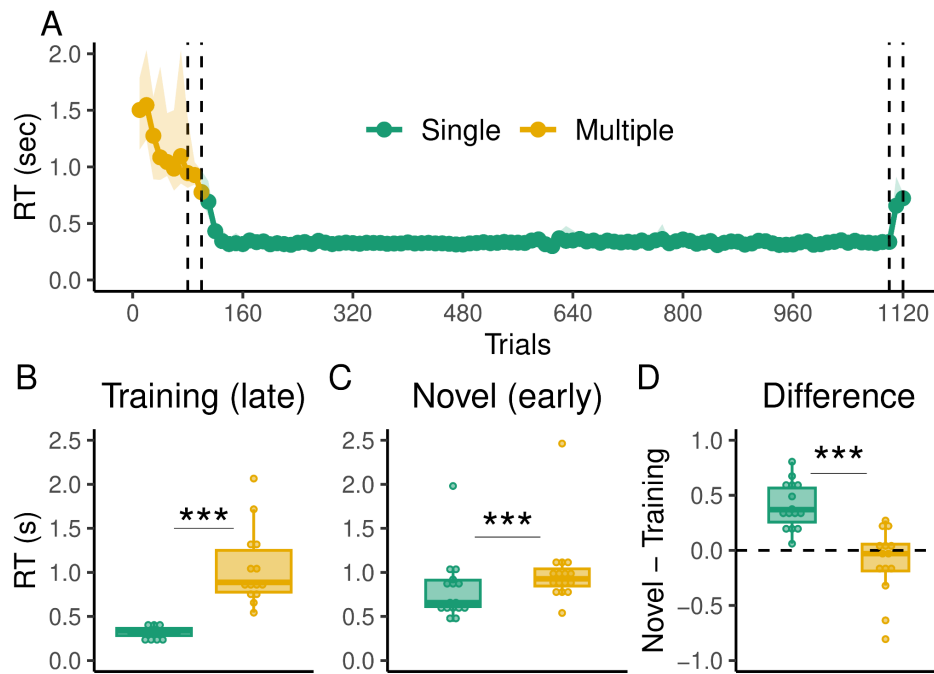


Figure 2.10: RTs in Experiment 3. **(A)** Per-subject medians of RTs across trial bins (1 bin = 10 trials). Gold data points indicate the Multiple trials and the green data points the Single trials. The first (trials 81-100) and second (trial 1100-1020) generalization phases are demarcated with the dashed vertical lines. The solid dotted line represents the median and the shading the interquartile range. **(B)** RTs in the last bin of training trials. **(C)** RTs in the first bin of generalization (novel pairs). **(D)** Difference in RTs between the first bin of generalization and the last one of training.

### Modeling results

We evaluated our previous models with the only change that, instead of the 2W hybrid model, we incorporated a hybrid model with four weights (4W hybrid model), one for every stage

of the experiment (Multiple trials, first generalization phase, Single trials and second generalization phase). Table 3 shows the results of the individual model comparison. We found that for both the AIC and BIC, the 4W hybrid model proved to be the best model for all participants. This finding was corroborated by our group model comparison, where there was a high probability (>99%) that a random subject taken from our population was best described by the 4W hybrid model (Figure 2.11A). Similarly, the exceedance probability indicated that the 4W hybrid model was better than the other models with high probability (>99%). Among all participants, this model was able to capture a median of 75% of the explainable variability in the data (Figure 2.11 B).

| Models | AIC                  | AIC wins | BIC                  | BIC wins |
|--------|----------------------|----------|----------------------|----------|
| MF     | 8910 [6991, 10941]   | 0/16     | 8924 [7005, 10955]   | 0/16     |
| MB     | 14562 [12110, 15820] | 0/16     | 14569 [12117, 15827] | 0/16     |
| 1W     | 8788 [6980, 10882]   | 0/16     | 8816 [7007, 10911]   | 0/16     |
| 4W     | 8595 [6838, 10750]   | 16/16    | 8644 [6887, 10799]   | 16/16    |
| AR     | 8750 [7159, 10978]   | 0/16     | 8778 [7187, 11006]   | 0/16     |

Table 2.3: Individual model comparison for Experiment 3. The AIC and BIC columns show the median across subjects and the interquartile range inside the square brackets. AIC wins and BIC wins columns indicate the number of participants where the given model was the best one according to each metric.

We compared the model-based weights of the hybrid model across the different experimental stages using a 2-Way ANOVA with experimental phase (training or generalization) and trial type (Single or Multiple) as factors, which revealed a significant main effect of the experimental phase ( $F(1,60) = 48.96$ ,  $p < 0.001$ ,  $\eta^2 = 0.31$ ) and trial type ( $F(1,60) = 23.36$ ,  $p < 0.001$ ,  $\eta^2 = 0.15$ ), indicating that the model-based weights were in general higher during

generalization and for the Multiple trials. However, post hoc pairwise comparisons revealed that there was no difference between the model-based weights between the generalization phases (adjusted  $p = 0.99$ ), but there was a significant difference during the Single and Multiple trials (adjusted  $p < 0.001$ ), with the model-based weights reaching near zero values for the prolonged Single trials (Figure 2.11C).

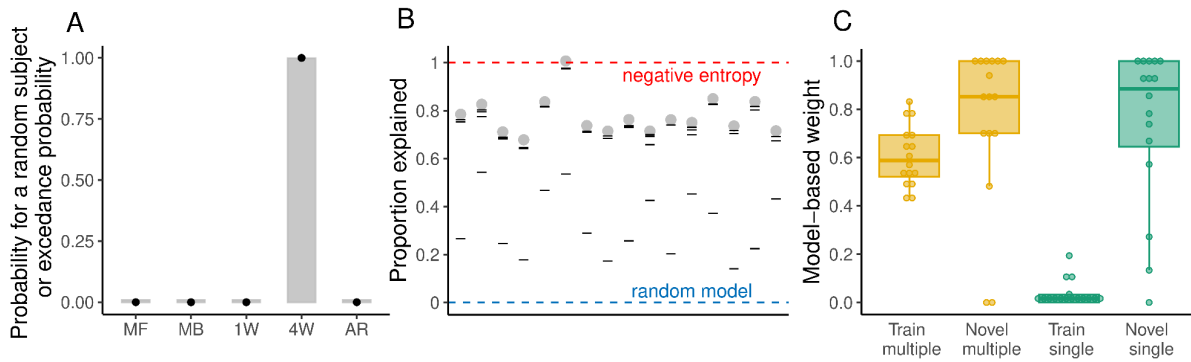


Figure 2.11: Modeling results of Experiment 3. **(A)** Probability that a random subject is best described by the tested models (MF: model-free, MB: model-based, 1W: Single-seight hybrid, 4W: Four-weight hybrid, AR: Arbitration model). Black dots indicate the exceedance probability. **(B)** Proportion of the variability in the data explained by the models. Gray dots represent this value for the best model according to the BIC and black lines represent the other models. Red and blue dashed lines represent the negative entropy (upper boundary) and the performance of a random model (lower boundary), respectively. **(C)** Model-based weight in the hybrid model during training and generalization in the Single and Multiple trials.

The results from Experiment 3 indicate that, even though participants were most likely relying on state-action associations during the prolonged period with no variability as reflected in low RTs, they were able to flexibly switch to model-based computations when presented with novel start-target pairs, as revealed by increased RTs and successful generalization. Such findings, which were supported by differences in the weights of the 4W hybrid model, show that a brief exposure to high variability can allow the formation of robust knowledge about the visuomotor mapping which can lead to successful generalization in the future.

## Experiment 4

It is well-established that skill learning can involve both explicit and implicit processes to varying degrees depending on the training conditions (Jiménez et al., 2006; McDougale et al., 2015; Taylor et al., 2014). In our prior experiments, the visuomotor mapping was relatively simple and deterministic (i.e., three keys map to three cursor directions) and, as a result, it is possible that participants may have developed explicit knowledge about it and/or the sequence of key responses. If successful generalization was dependent on an explicit representation of the mapping, one potential scenario is that generalization performance will be impaired, particularly in the Multiple group, when learning occurs implicitly. To address this possibility, we introduce a stochastic grid navigation task for the Single ( $n=16$ ) and Multiple ( $n = 16$ ) groups, where on every move there was a probability of 0.2 that the key press could move the cursor to any of the adjacent locations different from the original mapping (see *Materials and Methods* for details). Introducing stochasticity between stimulus-response mappings is a common method to blunt awareness and explicit learning in studies of motor sequence learning (Jiménez et al., 2006; Schvaneveldt and Gomez, 1998). In addition, we evaluated the level of explicit knowledge of the visuomotor mapping at the end of the task by asking participants where they thought the keys move the cursor to.

While adding stochasticity could impair the generalization performance in the Multiple group if the mapping was learned explicitly in our previous experiments, it is unclear the extent to which such variability in the key-outcome relationship would affect the performance in the Single group. One possible scenario is that, as in the Multiple group, any benefit from the explicit use of the mapping would go away, therefore making their generalization performance even worse. Alternatively, introducing key-outcome variability would, by default, prevent the repetitive sequence generation behavior that a deterministic mapping would induce, which could result in an enhanced use of the mapping.

Therefore, by leveraging this method, we ask two main questions: 1) Is better generalization performance observed in the Multiple condition, the result of an explicit representation of the visuomotor mapping? 2) Will the training variability induced by a stochastic mapping lead to better generalization in the Single group?

### ***Behavioral Results***

As a result of the stochasticity in the task, optimal arrivals were rare, therefore as a behavioral measure of performance we considered only arrivals to the target, regardless of the number of key presses. As in Experiments 1 and 2, the Multiple group learned more slowly than the Single group as revealed by a mixed effect analysis and where the Single group was significantly better than the Multiple group in 20 out of the 26 trial bins of training ( $p < 0.05$ ; Figure 2.12A). However, asymptotic performance was similar between the groups by the end of the training phase ( $t(26.73) = 1.31$ ,  $p = 0.19$ ; Figure 2.12B). More importantly, there were no differences in generalization performance ( $t(27.74) = -1.3351$ ,  $p = 0.1927$ ; Figure 2.12C) or in the change of performance from training to generalization ( $t(28.32) = -1.88$ ,  $p = 0.07$ ; Figure 2.12D). However, the Single group did significantly decrease its performance from training to generalization phases ( $t(15) = -2.65$ ,  $p = 0.01$ ) whereas the Multiple group did not ( $t(15) = -0.34$ ,  $p = 0.73$ ).

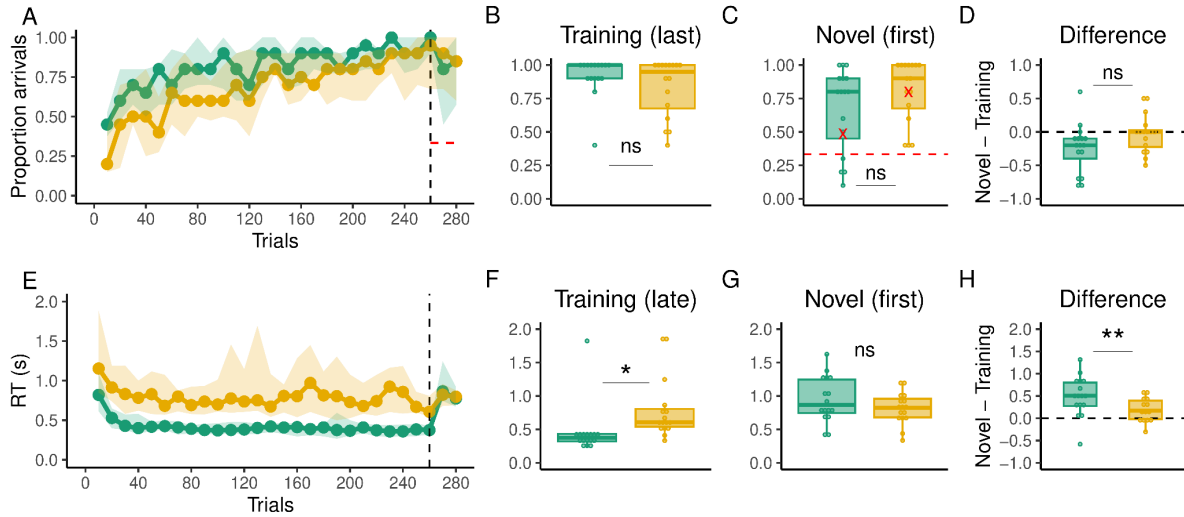


Figure 2.12: Behavioral results of Experiment 4. **(A)** Proportion of optimal arrivals over trial bins for the Single (green) and Multiple (gold) groups. The black dashed line indicates the beginning of the generalization phase. The solid dotted line represents the median and the shading region the interquartile range. The red dashed line indicates chance level of performance. **(B)** Proportion of optimal arrivals in the last bin of training trials. **(C)** Proportion of optimal arrivals in the first bin of generalization (novel pairs). Red marks indicate performance in the very first trial of generalization for all subjects. **(D)** Difference in the proportion of optimal arrivals between the first bin of generalization and the last one of training. The dashed line here indicates no performance change from training to generalization **(E)** RTs over trial bins. **(F)** RTs in the last bin of training trials. **(G)** RTs in the first bin of generalization **(H)** Difference in RTs between the first bin of generalization and the last one of training.

In contrast with the comparable performance in terms of target arrivals, we found that RTs at the end of the training phase remained significantly higher in the Multiple group ( $t(28.65) = -2.28$ ,  $p = 0.02$ ; Figure 2.12F). However, no difference in RTs was found between the groups at the beginning of the generalization phase ( $t(26.6) = 1.28$ ,  $p = 0.21$ ; Figure 2.12G) as the Single group significantly increased their RTs ( $t(15) = 11.05$ ,  $p < 0.001$ ; Figure 2.12H) to the level of the Multiple group. In contrast, no increase in RTs was found in the Multiple group ( $t(15) = 0.16$ ,  $p = 0.87$ ). Similarly, the increase in RTs was significantly greater in the Single group ( $t(29.36) = 2.74$ ,  $p = 0.009$ ; Figure 2.12H).

At the end of the experiment, participants performed an explicit test where they were asked to report the direction the key moved the cursor to. Both groups showed similar degrees

of explicit knowledge of the key-to-direction mapping (Figure 2.13A). For the Single group 68% (11 out of 16) knew all the keys correctly whereas 32% knew two, one or zero key directions. For the Multiple group 62% (10 out of 16) knew all the keys correctly while 38% (6 out of 16) knew two, one or zero key directions. We further explored whether participants that correctly knew the mapping (scoring 3 in the explicit test) in either group had better generalization performance than people that did not know it, or partially knew it (scoring lower than 3). While participants who had full knowledge of the mapping optimally arrived at the target more often on average than participants who had less knowledge, this difference was not significant ( $t(18.34) = -1.84$ ,  $p = 0.08$ ; Figure 2.13B). This suggests that explicit knowledge may not be a strong determinant of how people perform in the task.

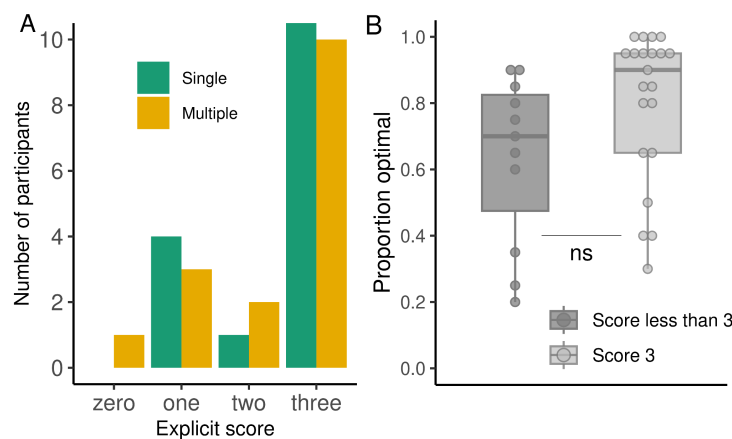


Figure 2.13: Results of the explicit test in Experiment 4. **(A)** Number of participants that correctly knew zero, one, two or three moving directions of the keys. **(B)** Proportion of optimal arrivals in the generalization phase for participants that score 3 or less than 3 in the explicit test of Experiment 4.

## Modeling Results

Table 4 shows the results of the model comparison at the individual level. We found that the 2W hybrid model best described the majority of the participants in both groups according to the AIC (12 out of 16 participants in the Single group and 13 out of 16 participants in the Multiple group) and BIC (9 out of 16 participants in the Single group and 8 out of 16 participants in the Multiple group) metrics. In addition, we found that there was a 94 % and 76% probability

that a random subject taken from the Single and Multiple groups, respectively, was best described by the 2W hybrid model (Figure 2.14A), and a > 99% probability that such model was better than the others overall according to the exceedance probability. Furthermore, we found that across participants the best performing model captured a median of 81% and 69% of the explainable variability in the data for the Single and Multiple groups, respectively (Figure 2.14B).

| Groups   | Models | AIC                | AIC wins | BIC                | BIC wins |
|----------|--------|--------------------|----------|--------------------|----------|
| Single   | MF     | 4096 [4297, 6157]  | 0/16     | 4702 [4309, 6170]  | 5/16     |
|          | MB     | 6066 [5458, 7233]  | 0/16     | 6072 [5464, 7239]  | 0/16     |
|          | 1W     | 4627 [4289, 6144]  | 3/16     | 4651 [4313, 6168]  | 1/16     |
|          | 2W     | 4620 [4276, 6137]  | 12/16    | 4650 [4306, 6168]  | 9/16     |
|          | AR     | 4631 [4310, 6144]  | 1/16     | 4655 [4334, 6169]  | 1/16     |
| Multiple | MF     | 7957 [5235, 10349] | 0/16     | 7970 [5247, 10362] | 0/16     |
|          | MB     | 8386 [5609, 10658] | 0/16     | 8392 [5615, 10665] | 0/16     |
|          | 1W     | 7752 [5032, 10133] | 1/16     | 7778 [5056, 10159] | 3/16     |
|          | 2W     | 7736 [5020, 10129] | 13/16    | 7768 [5050, 10162] | 8/16     |
|          | AR     | 7786 [5034, 10133] | 2/16     | 7812 [5058, 10159] | 5/16     |

Table 2.4: Individual model comparison for Experiment 4. The AIC and BIC columns show the median across subjects and the interquartile range inside the square brackets. AIC wins and BIC wins columns indicate the number of participants where the given model was the best one according to each metric.

We performed a 2-Way ANOVA over the model-based weights of the 2W hybrid model and found a significant main effect of the experimental phase ( $F(1,60) = 64.43$ ,  $p < 0.001$ ,  $\eta^2 = 0.47$ ) and the group ( $F(1,60) = 10.54$ ,  $p = 0.001$ ,  $\eta^2 = 0.07$ ), indicating that overall the weights were higher during generalization and for the Multiple group. Nevertheless, when performing post hoc pairwise comparisons, we found that, although there were clear numerical differences between the weights of the groups during training (Figure 2.14C), this difference did not reach significance levels (adjusted  $p = 0.07$ ). Similarly, there was no significant difference between the weights of the groups during generalization (adjusted  $p = 0.15$ ; Figure 2.14D).

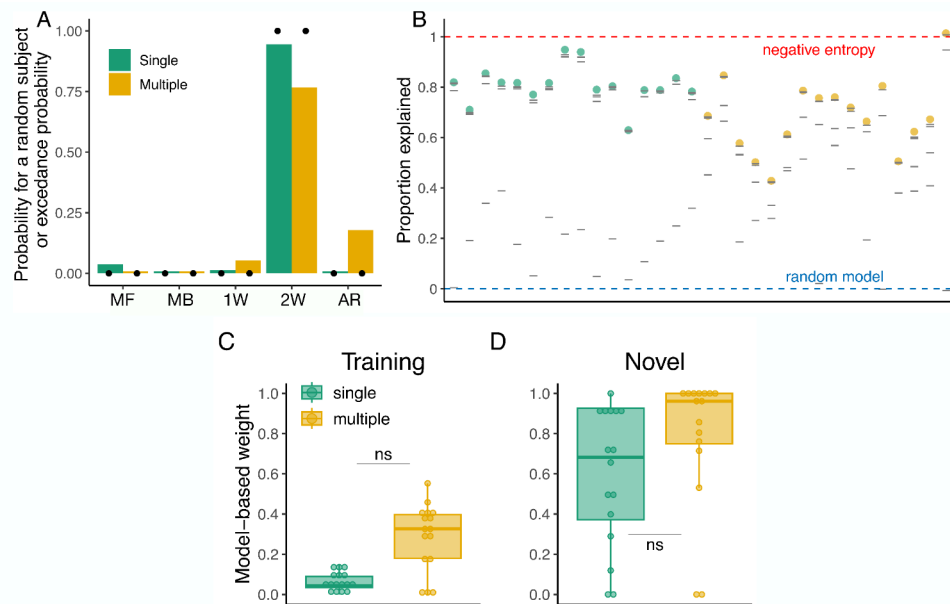


Figure 2.14: Modeling results of Experiment 4. **A:** Probability that a random subject taken from the Single and Multiple groups is best described by the tested models (MF: model-free, MB: model-based, 1W: Single-seight hybrid, 2W: Two-weight hybrid, AR: Arbitration model). Black dots indicate the exceedance probability. **B:** Proportion of the variability in the data explained by the models. Colored dots represent this value for the best model according to the BIC and gray lines represent the other models. Red and blue dashed lines represent the negative entropy (upper boundary) and the performance of a random model (lower boundary), respectively. **(C)** Model-based weight in the hybrid model during training and **(D)** generalization.

Overall, the results of Experiment 4 show that both groups had similar explicit knowledge about the mapping and that this knowledge was not related to participants' generalization performance. In addition, we found that adding stochasticity to the task improves generalization performance in the Single group to the level of the Multiple 2), as reflected in significantly lower RTs. Therefore, these results suggest that a group, suggesting that the former learned the mapping, potentially by preventing them from memorizing the sequence solution to the goal. While the Single group was able to generalize to levels comparable to the Multiple group, they did not show the computational demands of the latter during training (also observed in Experiment 1 and stochastic training with no variability in the start-target pairs, could bring the generalization benefits of training with multiple start-target pairs, while at the same time reducing its computational cost. Our modeling analyses indicate that such results could arise with stochastic training allowing to appropriately increase the weight of model-based computations during the generalization phase.

## **Discussion**

A vast number of skills require the formation of novel visuomotor mappings. Sometimes these mappings can be completely arbitrary like in video games, where an “up” press on a video game controller can lead a virtual character to move or jump. The advantage of learning and using these mappings, as opposed to simple state-action associations, is that the mappings can be used for planning and generalization to novel contexts (Tolman, 1948; Behrens, 2018; Kahn et al., 2018).

Across four experiments in a grid navigation task, we found that increasing the variability in the number of start-target pairs enables participants to more effectively use the visuomotor mapping and generalize to novel pairs. In particular, the Multiple group showed significantly better generalization performance than the Single group to pairs that required planning (Experiment 1), but also for the ones requiring no planning (Experiment 2). These results suggest that the lack of variability during training can impair the use of the visuomotor mapping

even for simple decisions. While the effects of training variability were clear during the generalization phase, we also found evidence that participants in the Multiple group made choices that reflected a greater use of the visuomotor mapping during training (Figure 2.3). In Experiments 3 and 4, we found that the limited generalization observed in the Single group can be enhanced by a short period of variability introduced early in learning or by incorporating stochasticity into the visuomotor mapping. In addition, our modeling results, along with differences in reaction times, indicate that the Multiple group assigned greater model-based control during the task, while the Single group struggled to switch from a model-free system to model-based computations when encountering novel pairs.

It is well known that training variability leads to better generalization performance in a variety of cognitive domains (Raviv et al., 2022). However, the evidence for such effects in motor learning tasks, particularly for motor adaptation, is mixed (Berniker et al., 2014; Braun et al., 2010). For example, in a visuomotor adaptation task, Berniker et al (2014) found that two groups of participants trained in workspaces of different sizes, which induced low or high variability in the reaching movements, had no differences in generalization performance on a novel workspace. In line with these results, several studies in motor adaptation suggest that learning proceeds rather locally (Mattar and Ostry, 2016; Berniker et al., 2013), with a very limited capacity to generalize beyond the trained range. In addition, Thoroughman and Taylor (2005) found that when the spatial complexity of a force field was increased, which necessarily increases variability, generalization actually narrowed rather than broadened.

On the other hand, Braun et al.c found that participants trained with random rotations, which would average out generating no learning but induce movement variability, were able to adapt faster to subsequent rotations compared to a control group. Their results suggest that training variability can lead to structural learning, i.e., visuomotor rotations, with subsequent learning primarily recalibrating the parameters of the rotation, i.e., the angle. It is important to note, however, that in subsequent work we found that the variability-induced benefit to

generalization was the result of explicit strategies and not implicit adaptation (Bond and Taylor, 2017).

Motor adaptation studies reporting limited generalization would suggest that learning in our task proceeds in a different manner (McDoughle et al., 2022). In fact, McDoughle et al. (2022) found that LSJ, a patient with severe bilateral hippocampal damage, was able to improve her performance in visuomotor adaptation tasks but was unable to learn a novel key mapping for grid navigation. Such results suggest that successful performance in the grid navigation task depends on brain regions known to be crucial for spatial navigation and planning (Mattar and Lengyel, 2022), in contrast to the cerebellar-dependant mechanisms involved in motor adaptation (Shadmehr et al., 2010; Cullen and Brooks, 2015; Sokolov et al., 2017).

It is possible, however, that some form of structural learning about the key sequences could have occurred during the training phase of Experiment 1, similar to the results of Braun et al. (2010), which could have resulted in better generalization for the Multiple group. Specifically, in Experiment 1, both the training and generalization phases had start-target pairs that were seven moves apart. However, by design, the Multiple group sampled a larger number of sequences during training, which may have inadvertently exposed them to elements of the trajectories needed to reach the targets in generalization. Although no start-target pair in generalization required the exact key press sequence as those in training, some prototypical movement patterns might have been reused to reach the novel targets. However, this scenario was ruled out in Experiment 2, where targets were only one move away during generalization, yet the Single group still showed significantly lower performance throughout the generalization phase. Therefore, we believe that differences in generalization performance observed between the Single and Multiple group were less likely to occur due to differences in structural learning, but instead due to differences in their ability to flexibly use their internal model (visuomotor mapping).

Notably, in spite of the differences in generalization performance observed in Experiment 1 and 2, the Single group did show evidence of using the visuomotor mapping, albeit not as effectively as the Multiple group. For example, they performed greater than chance when the novel targets were one move away. In addition, when contrasting the generalization performance in Experiments 1 and 2, their RTs increased more for distant targets than for proximal targets, indicating that they were planning (Redish, 2016; Jensen and Mattar, 2024). This result rules out the possibility that the observed change in RTs occurred solely due to the novelty of the targets, or that participants were just responding randomly. In addition, we also found that the inter-key-intervals during the generalization phase of Experiment 1, were significantly higher in the Single group, strongly suggesting that they had to revisit the visuomotor mapping to plan at intermediate steps of the trajectory. In terms of the modeling results, the Single group also showed model-based weights greater than zero during the training and generalization phases of Experiment 1 and 2, suggesting some influence of model-based computations. Lastly, even in the stochastic version of the task of Experiment 4, most participants in the Single group explicitly knew the moving directions of the keys to some extent, which makes it unlikely that they did not know them in the deterministic, and arguably simpler version of the task from Experiment 1 and 2.

We believe the inability of the Single group to use the visuomotor mapping as flexibly as the Multiple group was, in part, the result of the increased competition from habitual responses developed during the training phase. Indeed, a supplementary analysis from Experiment 1 and 2 revealed that, in the Single group more often than in the Multiple group, the errors during the generalization phase started with the same key as the one of the most frequently used sequence during training (Figure S2.6) – which suggest a persistence pattern characteristic of a habit. Our computational modeling analyses support these results as the model-based weights in the 2W hybrid model remained consistently lower overall in the Single group, but most importantly, following the transition from training to generalization, reflecting a greater influence

of model-free processes. Relatedly, when presented with stochasticity in the visuomotor mapping (Experiment 4), which in principle prevented participants from repeating the same sequence of key presses during training, the Single group matched the performance of the Multiple group in generalization. This type of training could have disrupted the formation of habitual responses and allow the flexible use of the learned mapping.

Previous research has demonstrated that a key characteristic of habitual responses is their persistence after reward devaluation (Adams, 1982; Dickinson, 1985). However, a history of reward might not be necessary for the formation of habits as they can emerge simply through the repetition of actions (Thorndike, 1911; Miller et al., 2019; Du et al., 2022). While our models implement reward as visually minimizing the (chessboard) distance to the target, it is unclear the extent to which this is actually rewarding for participants. In addition, the explicit presentation of rewards in our task (emoji faces at the end of the training trials) was relatively sparse. Therefore, it seems more plausible that any habitual response emerging during generalization, resulted from its frequent repetition during the training phase (Figure S2.6), rather than due to its strengthening from reward. One major limitation of our work is that the current version of our models do not explicitly account for such autocorrelation of responses. Instead, our implementation of the model-free system only predicts chance-level responses in the face of novel start-target pairs. This assumption is not accurate and would likely benefit from the incorporation of a persistence component of previously taken actions (Daw, 2011; Miller et al., 2018). The exact level at which persistence can occur in our task is an empirical question. For example, it can manifest at the global level if a subject has a tendency to press one key more often than the others throughout the task, regardless of the start-target pair or the current location of the cursor. Alternatively, persistence can occur at a more local level, where responses would be more likely to be repeated on a given grid state, or for a given start-target pair. Allowing the models to capture habitual responses during generalization could improve the

model fits and lead to a more pronounced difference in the model-based weights between the Single and Multiple groups.

In addition to the potential interference of habitual responses during generalization, the lower performance in the Single group could also be attributed to the abrupt transition from the training to the generalization phase, which the Multiple group experienced to a lesser extent given the trial-to-trial variability. While this is a plausible scenario in the very first generalization trial —where participants in the Single group did perform at chance level (Figure 2.6C) — it is unlikely to account for their consistently lower performance throughout the generalization phase. Although we did observe some recovery, the Single group continued to exhibit significantly lower performance than the Multiple group by the end of the generalization phase when we controlled for new learning (Figure 2.6A). In contrast, studies on change-point detection find that performance typically recovers immediately after an abrupt transition when a controller is well known, at least in simple scenarios (Nassar et al., 2010; Wilson et al., 2013).

Finally, a potential reason as to why the Single group could have exhibited worse generalization performance is because trajectory solutions for the unique start-target pair required the predominant use of the left-up and left-down keys, as compared to the right key. This asymmetry in responses could have led to greater errors associated with the latter key. We addressed this potential confound by analyzing if errors during generalization in the Single group were more frequently associated with the key moving the cursor to the right. We attempted to address this question by examining the generalization trials of Experiment 2 as they require a single move. We found that only 17% of the errors in generalization across all participants were related to this key. Therefore, lower generalization performance could not be attributed to the fewer training samples with the right key in the Single group.

In terms of our computational models, we leveraged a well known distinction in reinforcement learning between model-free and model-based algorithms (Daw et al., 2005; Botvinick et al., 2019). Although there is evidence that humans may use both systems (Gläscher

et al., 2010; Schultz et al., 1997), it is less clear the way in which they interact with one another. Similar to previous research (Daw et al., 2011; Gläscher et al., 2010), we assumed this interaction could occur as a simple weighted sum of the output of the systems with fixed weights. However, we also proposed a novel arbitration mechanism based on the encounters with familiar states. While this model did not provide the best fit at the population level, it provided the best fit for several participants at the individual level for Experiments 1, 2 and 4. It remains to be tested whether other arbitration models (Lee et al., 2014; Kim et al., 2019) can provide a better fit than our arbitration model or than the fixed weight hybrid models.

Similarly, further variations of the planning algorithm can be incorporated into the models. For example, even though Breadth First Search is suitable to solve deterministic tasks like ours, algorithms that incorporate heuristics, uncertainty or resource rational computations like pruning or truncation, seem more likely to be implemented by the brain (Mattar and Lengyel, 2022; van Opheusden et al., 2023). We decided to implement Breadth First Search (BFS) given its uniform structure which allows it to explore all paths that lead to the target (even if not immediately obvious). This was particularly important because participants often chose moves that deviated from the target but were optimal, which would have been harder to detect using search algorithms that typically favor moves toward the target (e.g., based on chessboard distance or Euclidean distance heuristics). While our model-free algorithm does tend to prefer such moves, the model-based algorithm aimed to capture the less intuitive moves that likely reflected planning. However, our implementation of BFS assumes that participants can plan with infinite depth, which is an unrealistic assumption. Future variations of this model can include limits in the planning depth known as “lookahead” which can be more biologically plausible (Mattar and Lengyel, 2022). Alternatively, learning algorithms that do not incorporate model-based computations have also been shown to generalize by transferring value across states and actions (Colas et al., 2022). However, the fact that reaction times were significantly higher in the Single and Multiple group for distant targets (Experiment 1) compared to proximal

targets (Experiment 2) strongly suggests that participants were engaged in planning in our task. Additionally, RTs in all our experiments were consistently higher than the inter-key intervals (Figure S2.2), which likely indicates the time participants spent preparing the sequence of key presses to reach the target—i.e., planning.

For Experiment 3, we found that the benefit of having variable training over generalization remained even after a long exposure to no variability. We believe this is a result of the formation of the novel visuomotor mapping at an early stage of learning (Fitts and Posner, 1967), and the abrupt change in variability could have separated the mapping memory from future updates (Gershman et al., 2010; Heald et al., 2021), preventing it from being forgotten. Subsequently, during the period of no variability, novel state-action associations could have been formed. If separate memories for the visuomotor mapping and the state-action associations were formed, the latter could have potentially been evoked by reducing the preparation time during generalization (Hardwick et al., 2019). Crucially, unlike in Experiments 1 and 2, the increase in reaction times from the end of the Single trials to the beginning of the second generalization phase was also accompanied by successful generalization. This suggests that participants were able to effectively switch to model-based computations.

The results of Experiment 3 corroborate previous findings outside visuomotor adaptation that indicate that the benefits of variable training occur when variability is introduced early in learning as opposed to later (Raviv et al., 2023), but only when variability is not too high. In our experiments, variable training implied being exposed to four pairs of start-target locations, which were repeated at least 20 times each (early in learning Experiment 3) and up to 70 times (Multiple conditions in Experiments 1, 2 and 4), which we believe provided participants enough familiarization with each of them. Had they experienced more variability, for example by changing the start-target pairs every trial, performance could have been slower and the benefits in generalization could have arrived later.

In Experiment 4, we showed that variability in the form of a stochastic mapping can lead to comparable generalization performance between the Single and Multiple groups. However, the significantly lower RTs during the training phase by the Single group, indicate that the improvements in generalization do not come at the expense of higher RTs during training. Therefore, a stochastic training with a single goal, can prove to be an effective regime that maintains the benefits of both model-free and model-based computations. In addition, while we were not able to fully remove the awareness of the mapping in the task, we did not find evidence that differences in awareness between participants were related to generalization performance. We believe further investigation is necessary to test the role of explicit processes in the learning of visuomotor mappings as previous research suggests they might be crucial in tasks similar to ours (McDougle et al., 2022; Wise and Murray, 2000).

Whereas most of previous studies in sequence learning like SRT,  $m \times n$  tasks or discrete sequence production have allowed the study of externally generated sequences specified by the experimenter, there has been a recent interest in sequences that humans generate internally (Bera et al., 2021), which, by not being constrained, allow us to explore the planning processes that make humans arrive at given solutions to achieve goals. A model task in this direction has been grid navigation. Our work provides a step in this direction by further providing cognitive models of the processes that might generate these sequences: model-based mapping learning and state-action associations. We believe these types of tasks are good models of a variety of the activities that humans perform in their lives such as playing video-games, musical instruments or sports, where improvisation and self selection of actions is a common feature. However, it is important to note that in contrast to the complexity of the mappings that humans learn in their lives, the mapping that we tested in our studies was relatively simple. Therefore, there is a need to study mappings with greater complexity. In a recent work (Velazquez-Vargas and Taylor, 2023), we address this point by asking participants to navigate in the grid using a mapping based on the less intuitive rule of the ‘Knight’ from chess. This and similar settings can

allow the study of the processes, such as learning and planning, involved in the acquisition of complex skills

Finally, grid navigation as in the current experiments sits at the intersection of motor learning and spatial navigation where the interaction of procedural and declarative processes likely occurs. For example, previous studies have highlighted that participants with some impairment in declarative knowledge do not perform as well as controls in similar tasks (McDougle et al., 2022). Therefore, grid navigation could be used as a testbed for how declarative knowledge contributes to the acquisition of a motor skill. At the same time, it rests at the level of complexity where it is still tractable to build relatively simple cognitive models to explain human performance.

### Supplementary material

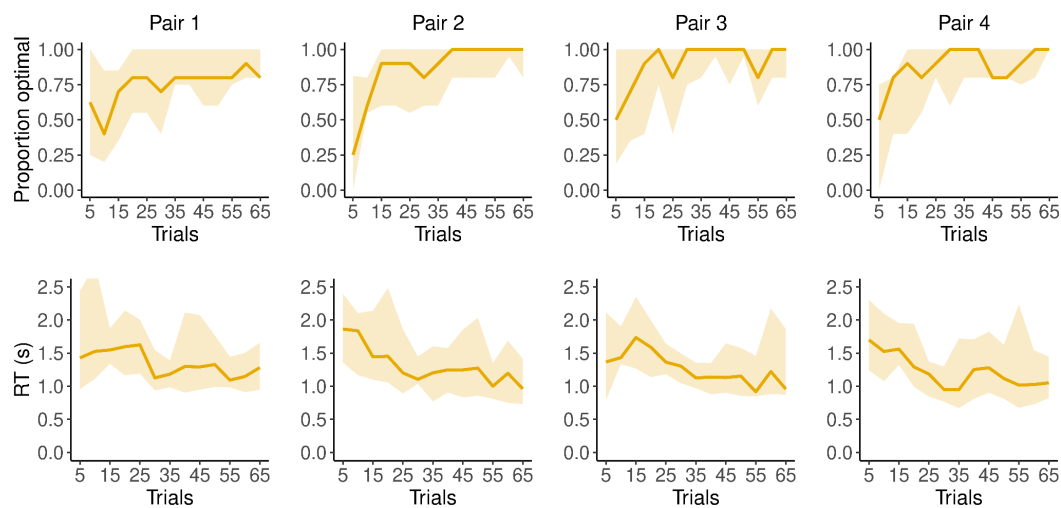


Figure S1: Target specific performance for the Multiple group in Experiment 1. No apparent differences were found in the performance between the different start-target pairs.

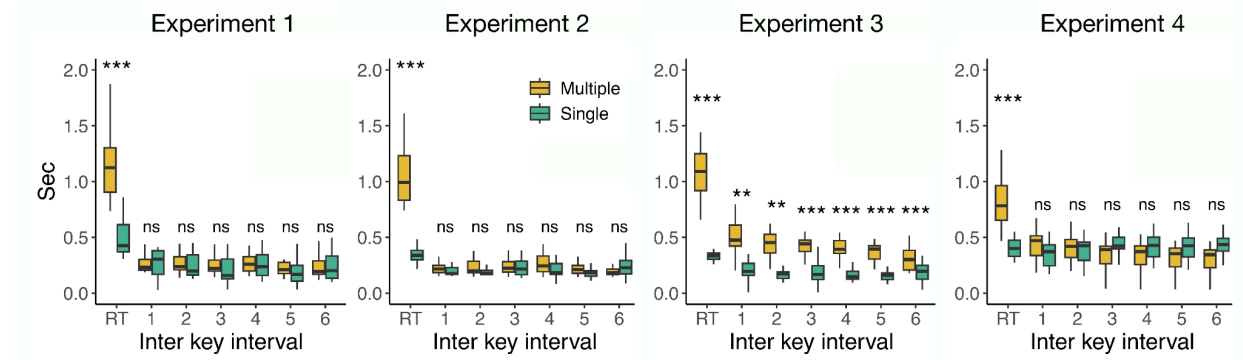


Figure S2.2: Reaction times and interkey intervals in the training phase of Experiments 1-4 (from left to right).

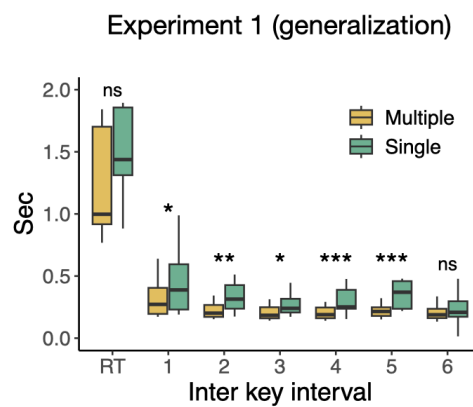


Figure S2.3: Reaction times and interkey intervals in the generalization phase of Experiments 1.

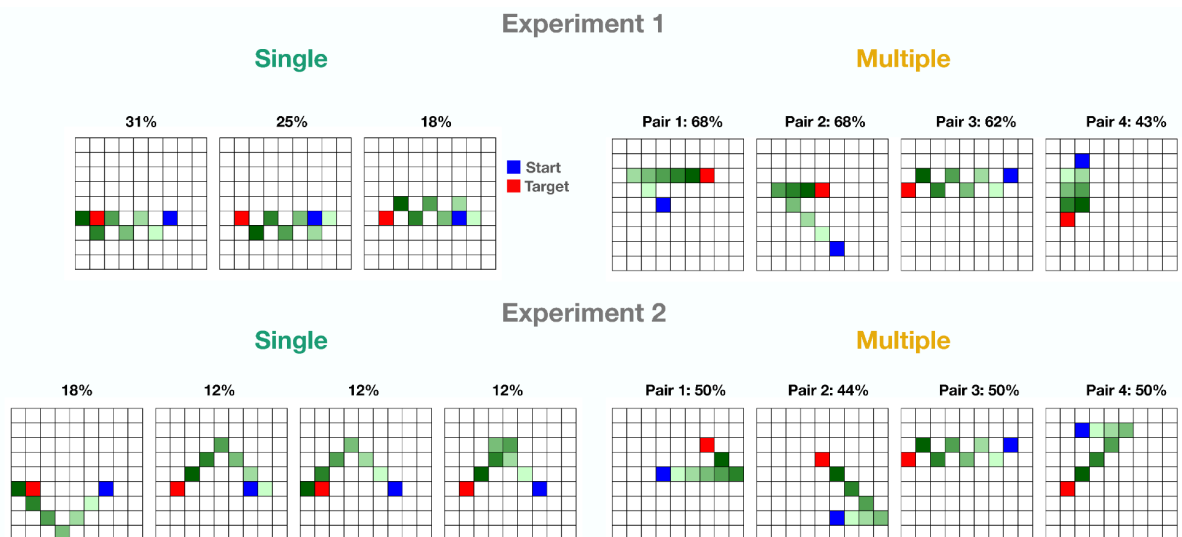


Figure S2.4: Most frequently used trajectories across participants in Experiment 1 and Experiment 2. Above each grid, we show the percentage of participants for which the depicted trajectory was the most frequently used during training. We show trajectories that were preferred by at least two subjects.

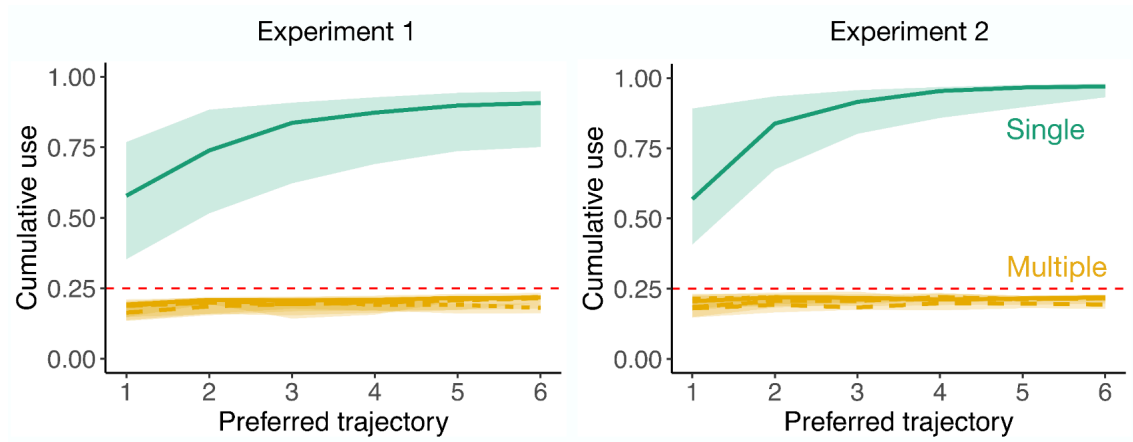


Figure S2.5: Cumulative use of the most frequent trajectories during the training phase in Experiment 1 and Experiment 2. The red line indicates the maximum value for the start-target pairs in the Multiple group (proportion = 0.25), given that each pair only appeared in one quarter of the trials. The different types of gold lines indicate each of the four start-target pairs in the Multiple group.

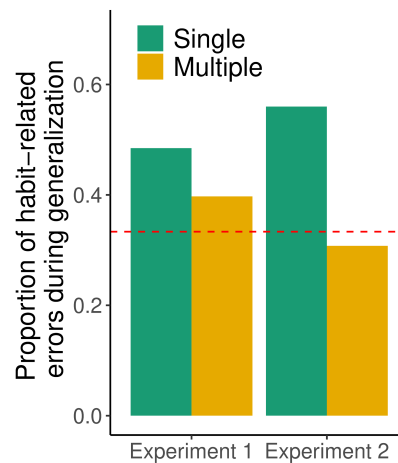


Figure S2.6: Evidence for the interference of habitual responses during the generalization phase in Experiment 1 and 2. For a given subject, we computed the number of errors during the generalization phase where the first move (the only move in Experiment 2) matched the first move of the most frequently used sequence during their training phase. Given that participants in the Multiple group did not make many errors, we aggregated these values across subjects and divided them by the total number of generalization errors across all participants. We report this value in the Y axis of the plot and the dashed redline denotes chance level.

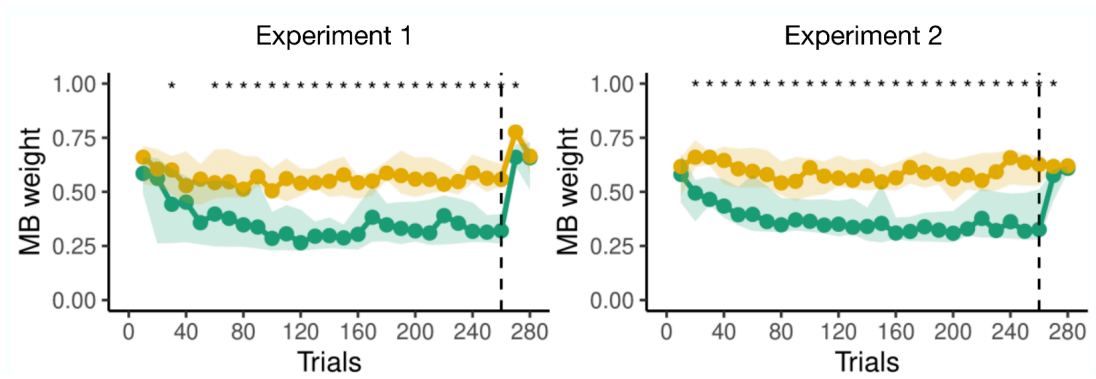


Figure S2.7: Model-based weight over trials for the unconstrained model in the Multiple (gold) and Single (green) groups of Experiment 1 and Experiment 2. The weights represent a time series of free parameters. A value of 1 reflects fully model-based, while a value of 0 reflects fully model-free. The dashed line demarcates the start of the generalization phase.

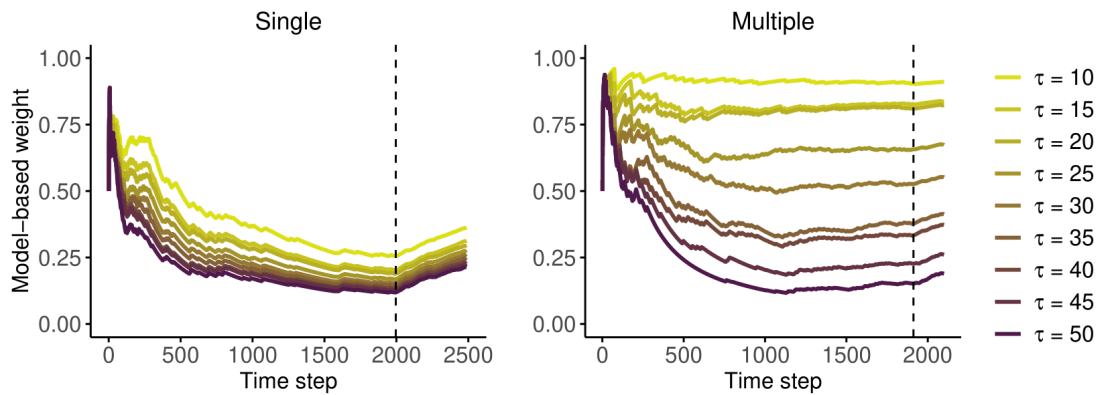


Figure S2.8: Simulations of the model-based weight over time in the arbitration model for different values of  $\tau$  in the Single and Multiple groups.

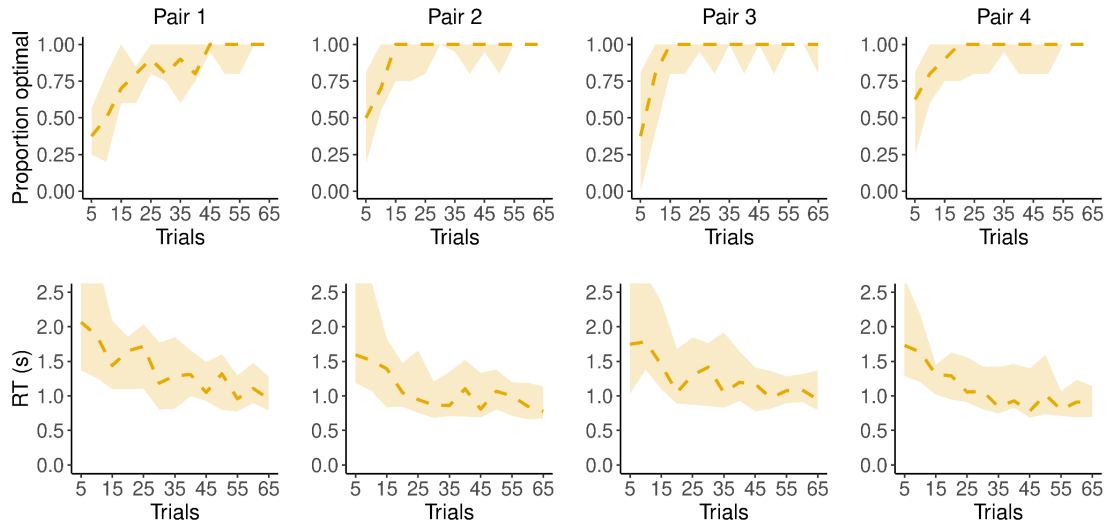


Figure S2.9: Target specific performance for the Multiple group in Experiment 2. No apparent differences were found in the performance between the different start-target pairs.

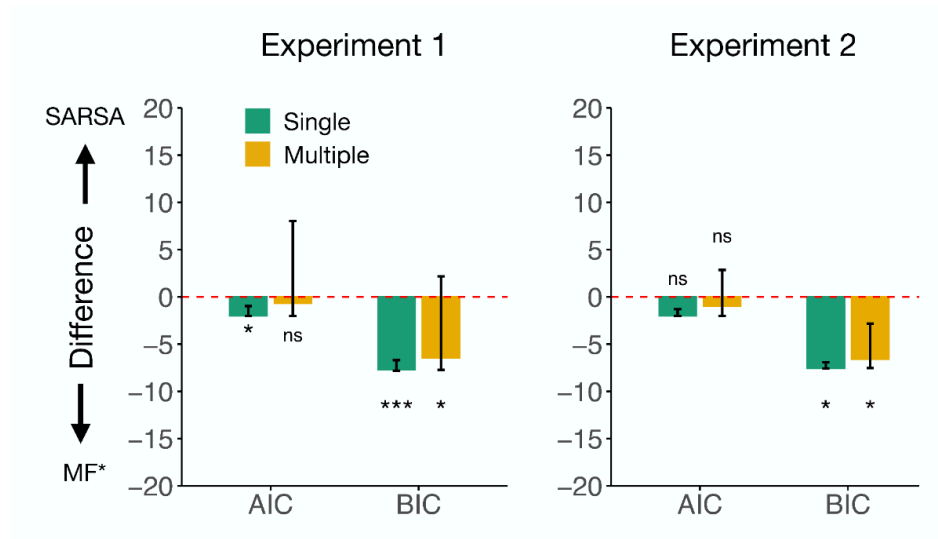


Figure S2.10: Model comparison between our model-free algorithm (here MF\*) and SARSA (20). SARSA provides a temporal difference update to state-action values for every start-target pair:  $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$ . We evaluated the models in the data of Experiment 1 and Experiment 2 using AIC and BIC differences and testing if they were different from zero using the Wilcoxon signed-rank test. We found that our model performed the same or better than SARSA depending on the metric. In particular, the AIC difference was significantly in favor of our model for the Single group ( $V = 34$ ,  $p = 0.04$ ), although there was no difference between the models in the Multiple

group ( $V = 77$ ,  $p = 0.33$ ). According to BIC, which penalizes more strongly the extra parameter  $\gamma$  in SARSA, our model-free algorithm was significantly better for the Single ( $V = 0$ ,  $p < 0.001$ ) and the Multiple group ( $V = 26$ ,  $p = 0.01$ ). We found similar results for the data on Experiment 2, with no significant difference in performance between the models for the Single ( $V = 46$ ,  $p = 0.13$ ) or the Multiple group ( $V = 67$ ,  $p = 0.97$ ) according to AIC, but our model outperformed SARSA for both the Single ( $V = 31$ ,  $p = 0.02$ ) and the Multiple group ( $V = 26$ ,  $p = 0.01$ ) according to BIC.

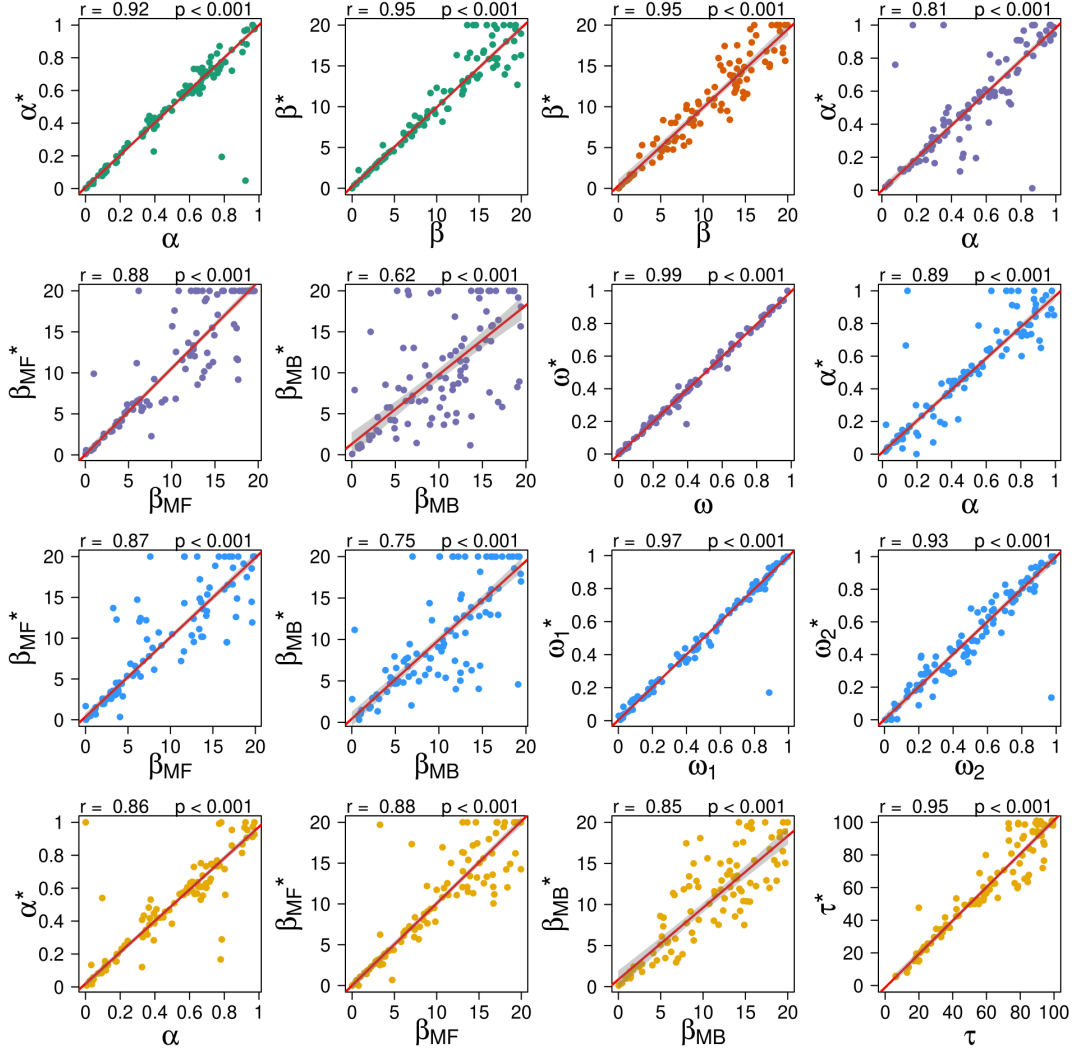


Figure S2.11: Results of parameter recovery. On the x axis is the simulated parameter and on the y axis the recovered parameter. The parameters for each model are indicated with different colors (MF = green, MB = orange, 1W = purple, 2W = blue, AR = gold):  $\alpha$  = learning rate,  $\beta$  = inverse temperature,  $\tau$  = memory window,  $\omega$  = model-based weight,  $\omega_1$  = model-based weight in training,  $\omega_2$  = model-based weight in generalization. Red lines represent the linear fit to the data and the gray shading the 95% confidence

interval. On the top of each plot we show the Pearson correlation between the simulated and recovered parameters as well as its associated p value.

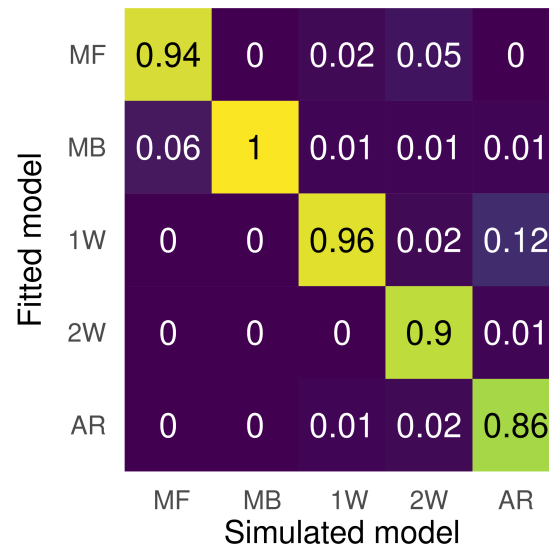


Figure S2.12: Confusion matrix with model recovery results. Numbers inside the cells represent the proportion of times that the model in the Y axis best recovered the data generated by the model on the X axis according to BIC.

## **Chapter 3**

### **Introduction**

In Chapter 2, I studied two different types of mappings that humans can learn when acquiring new motor skills. While the mapping participants learned provided valuable insights about its nature, it was relatively simple to acquire. Indeed, participants took around 80 trials (5-10 minutes) to reach asymptotic performance. This can contrast with more complex skills where people can take hours or even days to be learned. In this chapter, I address this point by making participants learn a more complex mapping, which also allows the study of the relationship between motor learning and planning more in detail.

Many of the skills that humans acquire in their lives require some extent of motor learning and planning. For example, when learning to play the guitar, one should learn which configuration of the hand generates a chord or a note. Subsequently, in order to generate a desired song, one must concatenate those configurations into sequences, a process which requires planning. The goal of this chapter is to study how skills requiring motor learning and planning are acquired.

Developments on motor learning and planning research currently provide no clear answers to this problem. 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 learned is fully specified by the experimenter (Korman et al., 2003; Kami et al., 1995). This limits considerably the planning aspect of the tasks, especially for internally-generated skills such as in musical improvisation or playing video games (Bera et al., 2021)

One exception is the work on grid-world navigation (Fermin, et al., 2010, 2016; Bera et al., 2021a, 2021b; Franklin and Frank, 2018, 2022; Velázquez-Vargas et al., 2024), where participants have to move a cursor from start to target locations using keys with an arbitrary movement-mapping. In these tasks, the learned mapping can be freely manipulated to generate sequences of actions to arrive at the targets, closely resembling video game playing. However, an important limitation of these studies is the small set of stimuli (start and target locations) that is commonly used, which makes it likely that participants readily use memorized solutions rather than performing planning (Velázquez-Vargas et al., 2024). Additionally, given that the mappings are typically simple, learning proceeds rather quickly, reaching ceiling performance within a few trials (Fermin et al., 2010, Bera et al., 2021a; Velázquez-Vargas et al., 2024). This contrasts with the complexity of the motor skills that humans can acquire, where the mapping can take hours or years to be fully learned and where planning processes are constantly being deployed due to the changing goals.

On the other hand, planning research, which focuses on how biological and artificial agents maximize their future rewards, has progressed significantly through the study of multi-step tasks (Solway and Botvinick, 2015; Gläscher et al., 2010; Daw et al., 2011) and navigation tasks (Tolman, 1948; Redish, 2016; Mattar and Daw, 2018; Johnson and Redish, 2007). In humans, various algorithms inspired by artificial intelligence (AI) have been proposed as hypotheses for how planning occurs (van Opheusden et al., 2017, 2023; Élteto et al., 2023; Krusche et al., 2018). While the specific algorithms humans use remain unknown, it is likely that the brain, given its limited resources, implements heuristics to reduce computational costs of planning, such as pruning or truncating the search of future scenarios (for a review, see Mattar and Lengyel, 2022).

A key component of most planning algorithms is the existence of an internal model of the world, which evaluates candidate actions based on knowledge about the rewards and state transitions. This model can be known beforehand such as in board games (van Opheusden et

al., 2021) or learned through experience as in instrumental learning experiments (Glächer et al., 2010; Daw et al., 2011). In most planning tasks, however, the internal model is assumed to require minimal or no motor learning at all. For instance, participants in multi-step choice tasks may only need to click on the intended choice to transition to a different state. Similarly, animals tested in mazes rely on a well-calibrated mapping of their body movements to navigate and reach rewards. While this assumption is reasonable for various cognitive tasks such as playing board games or performing spatial navigation, it contrasts with how motor skills are developed, where a mapping between state transitions and motor commands must also be learned.

In the present work, we developed a grid navigation task to study planning when state transitions require motor learning, while addressing the aforementioned limitations. In our experiments, participants moved a cursor from start to target locations using eight keyboard keys. Each key moved the cursor in L-shaped directions, following the movement of the chess piece known as the “Knight” (Velázquez-Vargas and Taylor, 2023). 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.

In Experiment 1, we show that learning the motor mapping in the absence of planning leads to performance benefits on subsequent stages of the task, where sequential decision making is needed. In addition, in Experiment 2, we show that incorporating mapping-learning into the planning process allows us to explain differences in performance for distinct planning horizons.

## **Materials and Methods**

*Participants:* Seventy five undergraduate students (34 males, 39 females and 2 non-binary; mean age = 19.6, sd = 1.5) from Princeton University were recruited through the Psychology Subject Pool. The experiments were approved by the Institutional Review Board (IRB). All participants provided written informed consent before performing the experiment.

*Apparatus and task design:* The experiments were carried out in person using the same computer equipment for all participants. 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. 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 were asked to move a cursor in the form of a frog from start to target locations represented by clovers (Figure 3.1A). The cursor moved deterministically based on the movement rule of the piece of chess known as the Knight. The Knight moves in L-shape directions: either two squares in one direction (up, down, right, left) followed by one square perpendicular to that direction, or one square in one direction followed by two squares perpendicular to that direction. Participants were not informed about which key moved the cursor in a particular direction. On a given position of the grid, the locations where the cursor could move to were highlighted in color in order aid learning. From a pilot study we found out that without this shading participants' performance was significantly worse, although they could learn eventually. Subjects moved the cursor using eight keys of their keyboard (A,S,D,F,H,J,K,L), which were randomly associated with a unique Knight move at the beginning of the experiment.

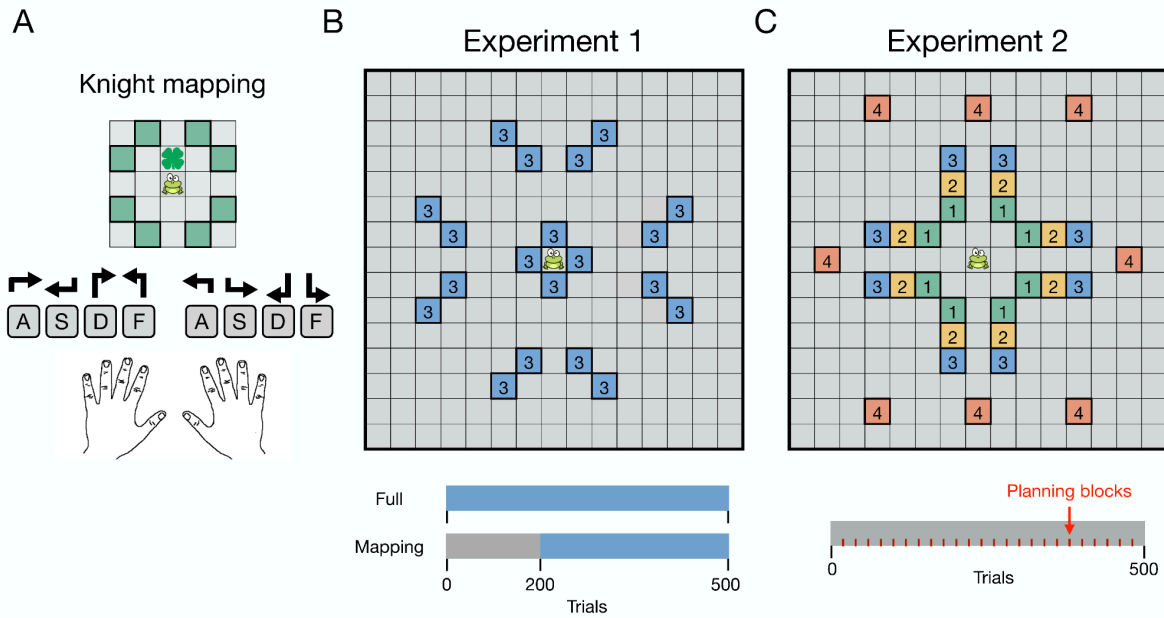


Figure 3.1. Grid navigation with the Knight mapping. **(A)** Participants were instructed to move a frog from start to target locations using eight keys, each corresponding to one of the Knight's possible moves in chess. The Knight moves in an L-shape directions: either two squares in one direction (up, down, right, left) followed by one square perpendicular to that direction, or one square in one direction followed by two squares perpendicular to that direction. To facilitate learning, the possible states the frog could move to at any given time were highlighted throughout the experiment. A single target location, out of 20 possible, was presented on every trial in the form of a clover **(B)** In Experiment 1, participants were grouped in two groups. In the Full group, they performed 500 trials with the targets being three moves away. Participants did not see the number indicating the distance to the target. In the Mapping group, participants experienced 200 trials with target locations being one move away in order to encourage the learning of the mapping. Subsequently, they performed 300 trials with the targets placed three moves away as in the Full group. **(C)** In Experiment 2, target locations were placed either one, two, three or four moves away from the start location. Every 20 trials, participants experienced a “planning block” consisting of four trials (one for each target distance) where no movement feedback was provided. During these trials, the frog turned blue, and participants were informed that it had frozen and it would not be able to move, but that they should still attempt to arrive at the target.

On every trial, the cursor appeared at the center of the grid and a target location was indicated with a clover. The distance to the target varied by experiment (see below). Importantly, only one pair of start-target locations appeared on 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. This points system has been used in previous grid navigation studies (Bera et al., 2021a). At the

end of each trial participants observed the points obtained as well as the cumulative points since the beginning of the experiment.

*Experiment 1:* The goal of this experiment was to test whether the acquisition of complex motor skills, involving sequential decision making, can be benefited from learning of the motor mapping first, followed by the incorporation of planning at a later stage. In order to tackle this question, participants performed the grid navigation experiment previously described in two groups. In the Full group ( $n = 25$ ), participants performed 500 trials and the target locations appeared always three steps away from the starting point (Figure 3.1B). We hypothesized that participants in this group face a double challenge: they need to arrive at target locations that require planning, and they need to do so while learning a non-trivial mapping. In order to reduce the likelihood that participants memorized the sequences to the targets and encourage planning, we used twenty different targets. They were presented in a random order, and each of them had to appear before showing them up again. The latter manipulation prevented the same target from being repeated multiple times. In the Mapping group ( $n = 25$ ), participants first experienced 200 trials with target locations being one move away in order to encourage the learning of the mapping. Participants were not informed that the first 200 trials had this purpose. Crucially, for this initial period, the trial was always terminated after the first move, such that participants could not generate further key presses to attempt to arrive at the target. Therefore, they could generate a maximum of 200 moves in the first 200 trials. Subsequently, they performed 300 trials with the targets placed three moves away as in the Full group.

*Experiment 2:* The goal of this experiment ( $n = 25$ ) was to directly assess planning processes throughout the course of learning a complex motor skill. In particular, we asked whether key signatures of planning such as differences in performance with varying distance to the targets (Jensen et al., 2024), would manifest in our task. In order to do so, we incorporated two main modifications. First, we manipulated the planning horizon by presenting targets that were either one to four steps away from the starting point. Participants experienced one trial

from each horizon at randomized order before experiencing a target with the same horizon again. The target locations were randomly chosen out of 32 unique locations (eight for each horizon; Figure 3.1C). In addition, every 20 trials we incorporated a “planning block” where no movement (state transitions) or performance (points) feedback was provided. In particular, participants were instructed that in some trials the frog would turn blue, indicating that it froze and would be unable to move. However, they should still make the responses attempting to reach the target location. These blocks consisted of four trials, one for each planning horizon, and were presented at a random order. The goal of the planning blocks was to encourage planning prior to generating the sequence as online corrections were not possible or significantly harder to make, i.e., based on the believed location of the frog. We hypothesized that this manipulation would maximize the effects generated by the different planning horizons. Importantly, unlike in the first 200 trials in the Mapping group of Experiment 1, the trials in Experiment 2 could only be terminated by arriving at the target or by exceeding the time limit (ten seconds). This was also true for targets that were one step from the starting point, which allowed participants to continue attempting to arrive at the target even if they missed in the first move.

**Computational models:** To provide further insight into the cognitive processes involved in acquiring complex motor skills requiring sequential decision making, we evaluated a set of computational models based on three components which have provided convincing evidence about their influence on behavior: planning (Hunt et al., 2021; Mattar and Lengyel, 2022; Tolman, 1948), reward learning (Sutton and Barto, 1998; Schultz et al., 1997), and action perseveration (Miller et al., 2019, 2021; Daw et al., 2011; Thorndike, 1911).

*Model-based planning (MB):* Due to the sequential nature of our task, we hypothesized that participants could estimate the value of their actions based on simulations about their future outcomes. In order to do so, we implemented a constrained version of A\* (“A-star”; Hart et al., 1968) as a planning model. A\* is a widely used search algorithm with various applications such

as in GPS navigation and in video-games (Rios and Chaimowicz, 2010). It is particularly well-suited for deterministic tasks like ours, offering significant computational advantages over uninformed algorithms such as Breadth-First Search (BFS; Moore, 1959) and Depth-First Search (DFS; Nilsson, 2014). These advantages come from A\*'s ability to efficiently balance path cost and heuristic information, leading to faster and more optimal solutions. Specifically, in A\* the cost  $f$  of a state  $s$  is defined by:

$$f(s) = g(s) + h(s) ,$$

where  $g(s)$  is the cost (number of steps) from the start state to the  $s$  state.  $h(s)$  is a heuristic function, which provides an estimate of the cost from state  $s$  to the target. In A\* search, paths are explored based on a priority queue where states with lower cost are prioritized.

As a heuristic function  $h(s)$ , we used the chessboard distance:

$$h(s) = \max(|x_s - x_{target}|, |y_s - y_{target}|),$$

where  $x$  and  $y$  represent the coordinates of the state  $s$  and the target. The chessboard distance provides an intuitive variable to guide planning in our task, as it simply involves counting the number of grid cells away from the target. Subsequently, we can calculate the cost  $C$  of transitioning to state  $s_i$  (where  $s_i$  represents any of the neighboring states to which the cursor can move to in the next step, following the Knight rule):

$$C(s_i) = f(s_i) + \sum_{j=1}^{D-1} f(s_j) \quad \text{for } D > 1.$$

Here, the first term represents the immediate cost of moving to  $s_i$ , while the second term represents the cumulative cost of the path starting from  $s_i$  as determined by A\* search.  $D$  is the depth of search, and is meant to capture the finite number of steps that humans plan into the future. In a model that considers only the next step  $C(s_i) = f(s_i)$ . For Experiment 1, we set  $D = 3$ , which corresponds to the initial distance to the target, as in this experiment we were

mostly interested in the learning of the transition probabilities and previous work has shown that this value falls within the range of the number of steps humans plan into the future (Huys et al., 2015; van Opheusden, 2023; Arad and Rubinstein, 2012; Snider et al., 2015). However, in Experiment 2, where we evaluated planning directly by manipulating the planning horizon, we compared models with different values of  $D$ .

Notably  $C(s_i)$  only provides the cost of moving to the available states, but not how to do it. The mapping between the keys and the state transitions still needs to be learned. In previous work (Velázquez-Vargas et al., 2024), we proposed that this learning process can be modeled using Bayes' rule. In this approach, each state transition is treated as a category, with new observations updating the beliefs about the true transitions through a Dirichlet-categorical model. However, an important limitation of this approach is that it does not account for differences in the learning rates of the state transitions, whether by individuals or due to experimental conditions. To address this, we adopted an alternative update rule based on prior work on reinforcement learning (Lee et al., 2014). Here, the transition probability from state  $s$  to state  $s_i$  after taking action  $a_k$  is given by:

$$T(s, a_k, s_i) = T(s, a_k, s_i) + \alpha_{state} (A - T(s, a_k, s_i)),$$

where  $\alpha_{state}$  is the learning rate for the state transitions.  $A = 1$  if action  $a_k$  is chosen and  $A = 0$ , otherwise. Importantly, in our task  $T(s, a_k, s_i)$  is the same for all  $s$ , as the mapping does not change depending on the location on the grid.  $T(s, a_k, s_i)$  was initialized as  $\frac{1}{8}$ , indicating that the eight transitions from the Knight rule were equally likely for all keys at the beginning of the experiment. Subsequently, the cost of action  $a_k$  is computed as:

$$C(a_k) = \sum_{i=1}^8 T(s, a_k, s_i) C(s_i).$$

Finally, the probability  $\phi$  of selecting action  $a_k$  is generated using a Softmax function and

–  $C(a_k)$ , so actions with less cost are more valuable:

$$\phi_k = \frac{e^{-\beta C(a_k)}}{\sum_{i=1}^8 e^{-\beta C(a_i)}},$$

where  $\beta$  is the inverse temperature parameter. This model has two free parameters:

$\alpha_{state}$  and  $\beta$ .

*Model-free learning (MF)*: This model assumes that participants update the value of their actions using reward prediction errors (Sutton and Barto, 1998; Schultz et al., 1997; Rescorla and Wagner, 1972; Niv and Schoenbaum, 2008). In particular, we implemented SARSA (Sutton and Barto, 1998), a common temporal difference model:

$$Q(s, a) = Q(s, a) + \alpha_{reward} (r + \gamma Q(s', a') - Q(s, a)),$$

where  $Q(s, a)$  is the current value of the state-action pair,  $\alpha_{reward}$  is the learning rate and  $r$  the rewards, which corresponded to the actual points at the end of the experiment.  $\gamma$  is a discount factor and  $Q(s', a')$  the estimated value of the next state-action pair. Importantly, in order to reduce the computational demand of storing state-action pairs for each target location, we used a state representation based on relative coordinates with respect to the target. This way, the state-action values can be reused across different targets. Lastly, the probability  $\phi$  of selecting an action is generated using a Softmax function:

$$\phi_k = \frac{e^{\beta Q(s, a_k)}}{\sum_{i=1}^8 e^{\beta Q(s, a_i)}}.$$

This model has three free parameters:  $\alpha_{reward}$ ,  $\gamma$  and  $\beta$ .

*Action perseveration (Persev)*: Previous studies have shown that actions can be strengthened merely through repetition, independent of the reward obtained (Miller et al., 2019,

2021; Daw et al., 2011; Thorndike, 1911), leading to the formation of habits. Given the relatively infrequent rewards compared to the number of key presses in our task, we hypothesized that the perseveration on previous actions could be a contributing factor on participants' choices. In particular, we implemented an update rule based on Miller et al., (2021):

$$H(s, a) = H(s, a) + \alpha_{persev}(r - H(s, a)),$$

where  $H(s, a)$  is the perseveration value of the state-action pair,  $\alpha_{persev}$  has a similar function of a learning rate, and determines the influence of past choices, with values closer to 1 indicating a greater weight of the most recent choices. Here,  $r = 1$  if the action for the given state-action pair is selected, and  $r = 0$ , otherwise. Similar to the MF algorithm, we adopted a state representation based on relative coordinates. The probability  $\phi$  of selecting an action is then generated using a Softmax function:

$$\phi_k = \frac{e^{\beta H(s, a_k)}}{\sum_{i=1}^8 e^{\beta H(s, a_i)}}.$$

This model has two free parameters:  $\alpha_{persev}$  and  $\beta$ .

*Hybrid models:* We considered the possibility that each of the above-mentioned components could influence participants' behavior. Specifically, we propose that responses could be the result of a linear combination of the MB, MF and perseverative components. We evaluated three hybrid models with two components and the full model with the three components. In the three-component hybrid, the value  $U$  of actions on a given state is given by:

$$U = \omega_{MB}C + \omega_{MF}Q + \omega_{persev}H$$

where  $\omega_{MB}$ ,  $\omega_{MF}$ ,  $\omega_{persev}$  are free parameters with values between [-1,1]. In order to have values of C, Q and H that were within the same range, we normalized them to be between [0,1]. For the two-component hybrid models, one of the three terms was omitted. The probability  $\phi$  of selecting an action is generated using a Softmax function:

$$\phi_k = \frac{e^{\beta U_k}}{\sum_{i=1}^8 e^{\beta U_i}}.$$

The number of free parameters in the hybrid models ranges between 5-8 depending on the components incorporated.

**Behavior analyses:** In Experiment 1 and 2, we used the points obtained in the task over trials as a measure of performance. In addition, in Experiment 1, in order to control for the difference in the number of key presses between the groups, we also computed the points by key presses. To do so, we grouped the key presses in bins of 40, and for each subject we obtained the averaged points they obtained in that period. Importantly, only the last responses in a trial can earn points. For example, if in 40 key presses a subject arrived at the target 10 times, the score for the bin corresponded to the average of the points obtained in those 10 times.

**Model analyses:** The parameters from our models were estimated at the individual level using Bayesian Adaptive Search (Acerbi and Ma, 2017) implemented in Matlab code. We verified that the model parameters of the best fitting model (MB-Persev) were recovered reliably (Figure S3.1), and that the tested models with different depths were identifiable among each other (Figure S3.2). For Experiment 2, model fitting was based on the trials with feedback as in the planning blocks the states were no longer observable, and it is unknown where the updates from learning would be attributed to. Simulations on Figure 3.6 were performed by randomly sampling 100 target locations from the ones that participants could encounter. In addition, we used the mean of the best fitting parameters from the Full group in Experiment 1, with the exception of the inverse temperature parameter which we set to  $\beta = 15$ , to reduce exploration. The values estimated for  $\beta$  from the Full group were near 1, however, the model was fit on responses while we were trying to predict the outcome of the entire sequence (trial points). Simulations on Figure 3.7 were generated in the same way but using the best fitting parameters in Experiment 2.

## Results

### Experiment 1

Acquiring new motor skills can be challenging since they often require simultaneous learning of the motor mapping and planning. In this experiment, we asked whether separating the processes of learning and planning can speed skill acquisition. To do so, two groups of participants performed a grid navigation task where the goal was to move a cursor from start to target locations using eight keyboard keys, which were mapped to one of eight directions according to the Knight rule in chess (Figure 3.1A). If participants arrived at the target using the minimum number of key presses, they received 100 points and lost 5 points for every extra move. In the Full group ( $n = 25$ ), participants performed the experiment while attempting to arrive at target locations that were always three steps away. We hypothesized that in this group, participants faced a double challenge: learning a complex mapping and performing sequential decision making. In contrast, the Mapping group ( $n = 25$ ), first experienced 200 trials with target locations placed one step away to encourage the learning of the mapping before having to plan. Importantly, during this period the trials were terminated after the first move such that participants could generate a maximum of 200 moves. Subsequently, participants performed 300 trials with the target locations being three steps away as in the Full group.

*Behavioral results:* To contrast the performance between the experimental groups, we first computed the points they obtained over trials (Figure 3.2A). During the first 200 trials, the Mapping group performed overall better ( $t(47.98) = 8.06$ ,  $p < 0.001$ ). This result is expected as the target locations for the Mapping group were only one step away. However, we did not find any significant differences in the overall performance between the groups after trial 200 ( $t(41.56) = 0.37$ ,  $p = 0.71$ ). These results could suggest that learning the mapping first did not provide any advantage in the task when the targets were three steps away compared to the Full group. It is important to note, however, that the comparison is not balanced. During the first 200 trials, the

Mapping group could generate a maximum of 200 key presses as the trials were terminated with the first key press. In contrast, participants in the Full group did not have this constraint. Indeed, they had a median of 7 key presses per trial during the first 200 trials, and therefore built considerably more experience than the Mapping group. In order to account for this, we grouped the points by key presses rather than by trial (see *Materials and Methods* for details; Figure 3.2B).

Similar to the comparison across trials, we found that the Mapping group performed better than the Full group for the first 200 key presses ( $t(35.42) = -14.75$ ,  $p < 0.001$ ). After key press 200, the performance of the Mapping group significantly dropped ( $t(39.57) = 11.57$ ,  $p < 0.001$ ) to the level of the Full group (first bin after key press 200;  $t(41.87) = -1.39$ ,  $p = 0.16$ ). However participants quickly began to recover their performance, and were significantly better than the Full group in all the following bins ( $p < 0.001$ ) as revealed by a mixed-effects model analysis. Therefore, when accounting for the number of key presses made in the task, the Mapping group performed remarkably better than the Full group with targets three steps away, suggesting an advantage from learning the mapping first.

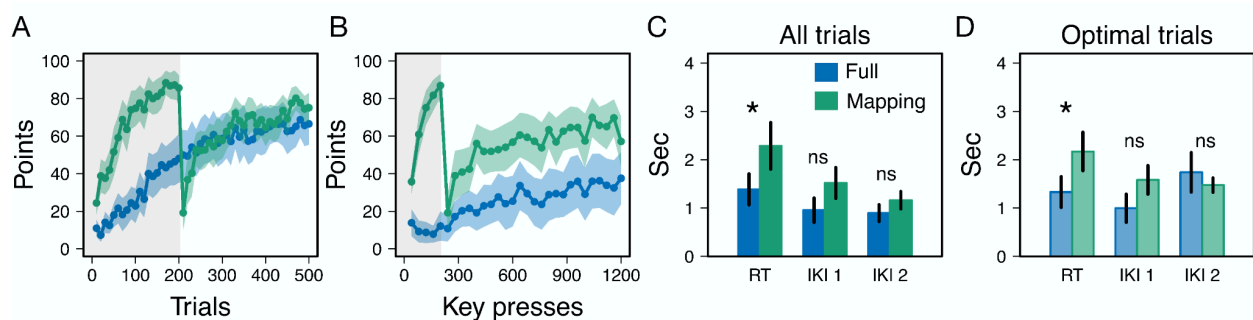


Figure 3.2: Behavioral results of Experiment 1. **(A)** Points over trials for both experimental groups. The shaded area indicates the first 200 trials (or key presses) where the targets were one step away for the Mapping group. **(B)** Points grouped by key presses. **(C)** RTs and IKIs across all trials. **(D)** Same as in (C) but for optimal trials.

To gain insights about the planning processes in our task, we analyzed the reaction time (RT) and inter-key intervals (IKI) when both experimental groups performed the task with the same complexity, i.e., when target locations were three steps away. For the Full group we analyzed the first 300 trials of the task, and for the Mapping group the 300 trials following the exposure to the targets one step away. We performed a 2X3 mixed-design ANOVA with the experimental group as between-subjects factor (Full and Mapping) and time interval as within-subjects factor (RT, first IKI 1 and second IKI). We found that the Mapping group took, in general, took longer to generate their responses as revealed by a significant main effect of the experimental group ( $F(1,48) = 9.75$ ,  $p = 0.003$ ,  $\eta_p^2 = 0.17$ ; Figure 3.2B). Pairwise comparison with Bonferroni correction revealed that RTs were significantly higher in the Mapping group ( $p = 0.03$ ) but we found no differences between the groups in the first ( $p = 0.09$ ) or second ( $p = 0.5$ ) IKI.

We conducted the same analysis for both RT and IKI, focusing exclusively on trials where participants reached the target optimally, i.e., using three moves. These trials were most representative of planning given that the probability of arriving at the target this way without already knowing the mapping is considerably low (see Table 3.1). Similar to the results for all trials, we found a significant main effect of the experimental group ( $F(1,48) = 7.85$ ,  $p = 0.007$ ,  $\eta_p^2 = 0.14$ ; Figure 3.2C), where the Mapping group generally had higher RTs. Following the same pairwise comparison method, we found that RTs were significantly higher in the Mapping group ( $p = 0.02$ ) but there were no differences in the first ( $p = 0.08$ ) or second IKI ( $p = 0.99$ ). In addition, this time we found that the second IKI in the Full group — which corresponds to the response prior to reaching the target — was significantly higher than the first IKI ( $p = 0.03$ ). We did not find this difference for the Mapping group ( $p = 0.99$ ). The latter result indicates that participants in the Full group potentially paused to re-plan the last move before reaching the target.

| Distance from the target | p(optimal)     | p(arrival)<br>Max. moves = 8 |
|--------------------------|----------------|------------------------------|
| 1 move                   | 0.125          | 0.2                          |
| 2 moves                  | 0.031          | 0.07                         |
| 3 moves                  | 0.011          | 0.03                         |
| 4 moves                  | [0.002, 0.005] | [0.011, 0.019]               |

Table 3.1. Probability of randomly reaching the target with the optimal number of moves (second column) and within eight moves (third column). These probabilities were computed using Breadth First Search (BFS). A maximum of eight moves was considered based on the observation that participants on average made 7.43 moves across both experiments

In summary, our behavioral results show that learning the mapping of the task first, can bring significant performance benefits in subsequent decisions involving planning. In addition, the response time analyses showed that participants in the Mapping group took significantly longer to initiate their responses, while the Full group paused before generating the last response for the trials most representative of planning. This last finding suggests that, while the Mapping group may have planned the whole action sequence beforehand, the Full group most likely relied more on intermediate feedback and replanned before the last move.

*Modeling results:* We evaluated seven computational models with the goal of explaining the differences in performance between the experimental groups (See *Materials and Methods* for details). These models were inspired by previous research demonstrating the influence of three main variables on decision making: planning (Hunt et al., 2021; Mattar and Lengyel, 2022; Tolman, 1948), reward learning (Sutton and Barto, 1998; Schultz et al., 1997), and action perseveration (Miller et al., 2019, 2021; Daw et al., 2011; Thorndike, 1911). We represented planning using a model-based (MB) algorithm which implements A\* search (Hart et al., 1968), and learns the motor mapping between the keys and the state transitions using state prediction

errors (Lee et al., 2014). For reward learning (MF), we implemented SARSA (Sutton and Barto, 1998), a widely used temporal difference model which updates state-action values based on the points obtained in the task. For action perseveration, we implemented a model based on Miller et al. (2021) where the value of actions depends on how frequent they have been repeated in the past. In addition, we consider four hybrid models represented as linear combinations of the aforementioned models.

Given that the MB algorithm captures our main variables of interest (the mapping learning and planning), we predicted that the differences in performance could manifest in elements of this model. In particular, we hypothesized that, given the faster improvement in the task by the Mapping group, they would show higher learning rates about the state transitions compared to the Full group.

We followed the method from Stephan et. al (2009) to obtain the probability that a subject selected at random from our experimental groups, was best described by the proposed models. This probability was 99% and 85% in the Full and Mapping groups, respectively, for a two-component hybrid model incorporating model-based planning and action perseveration (MB-Persev; Figure 3.3A). This model was able to capture on average 47% (Full group) and 57% (Mapping group) of the explainable variability in our data as compared to an information-theoretic near upper bound (Shen and Ma, 2016; Grassberger, 1988, 2003), where 0% represents chance level.

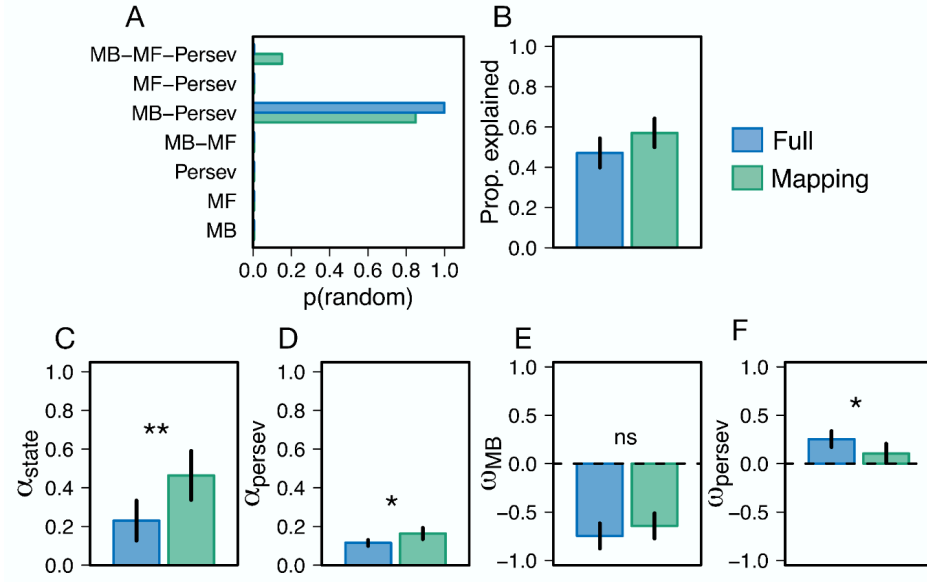


Figure 3.3: Model results of Experiment 1. **(A)** Probability of obtaining each of the models for a randomly selected participant in the Full and Mapping groups. **(B)** Proportion of the variability in the data explained by the best-fitting model (MB-Persev). **(C-F)** Individual parameter values in the MB-Persev for each of the experimental groups ( $\alpha_{state}$  = learning rate for the state transitions;  $\alpha_{persev}$  = learning rate for previous responses;  $\omega_{Persev}$  = weight of the perseveration component;  $\omega_{MB}$  = weight of the model-based component).

As predicted, we found that the Mapping group had significantly higher learning rates for the state transitions, i.e. the motor mapping ( $\alpha_{state}$  parameter;  $t(46.2) = -2.94$ ,  $p = 0.005$ ). In addition, we found that responses that were more distant in the past, had a greater influence for the Full group as indicated by significantly lower values of the parameter  $\alpha_{persev}$  ( $t(36.36) = -3.02$ ,  $p = 0.004$ ). Similarly, we found that the weight of the perseveration component in the hybrid model,  $\omega_{Persev}$ , was significantly higher for the Full group ( $t(46.37) = -2.35$ ,  $p = 0.02$ ). We did not find any significant differences in weight for the model-based component,  $\omega_{MB}$ , where its negative values indicate preference for options that incur overall less cost (see *Materials and Methods* for details). Briefly, the cost of an action is determined by the A\* search algorithm,

which estimates the cost of reaching the target if that action is taken, and this cost is adjusted based on the uncertainty of the mapping.

Overall, our modeling results indicate that better performance of the Mapping group when encountering targets that required sequential decisions, could be attributed to multiple factors. First, as shown by differences in the learning rates for state transitions, the initial training without planning in the Mapping group, could have allowed them to learn the mapping between the keys and the state transitions faster. This, in turn, would have facilitated planning in subsequent stages of the task. Furthermore, the Mapping group had a weaker influence from past responses (Figure 3.3D) and were less perseverative (Figure 3.3F) compared to the Full group. Therefore, it is likely that the Mapping group relied more heavily on planning processes rather than habitual responses, which is as also suggested by their higher RT (Figure 3.2C).

While the sequential nature of our task suggests that planning plays an important role for successful performance, Experiment 1 does not directly manipulate task variables to measure planning. In Experiment 2, we sought to investigate this more directly by varying the distance to the target (i.e., the planning horizon), which previous work has shown to affect the response time associated with the increased demands of planning (Jensen et al., 2024).

## **Experiment 2**

In this experiment, we directly assessed planning in our grid navigation task by manipulating the target distance from the cursor. We hypothesized that key signatures of planning, such as increased RT with the target distance (Jensen et al., 2024) would manifest, and that participants' performance would decrease for targets further apart due limitations in planning depth.

As in Experiment 1, participants performed 500 trials of the task. However, this time the target could appear from one to four steps away from the start location (Figure 3.1B). In addition, we incorporated "planning blocks" every twenty trials, where movement and

performance feedback was omitted (see *Materials and Methods* for details). The aim of these blocks was to prevent participants from relying on online path corrections, encouraging planning of the whole sequence prior to initiating responses. We hypothesized that this manipulation would maximize the effects generated by the different planning horizons. We refer to the trials in the planning block as “no-feedback trials” and to the trials with feedback as “feedback trials”.

*Behavioral results:* In order to assess participants’ performance in the task, we used a 4X2 repeated measures ANOVA, where we found a significant main effect of target distance (one to four steps away;  $F(3,72) = 60.28$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.71$ ) and trial type (feedback and no-feedback;  $F(1,24) = 61.75$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.72$ ) in the points obtained, which indicated that participants performed overall better in the feedback trials and at shorter target distances (Figure 3.4A). Post hoc pairwise comparisons (Bonferroni corrected), revealed that performance was better in the feedback than in the no-feedback trials for targets at two ( $p = 0.002$ ), three ( $p < 0.001$ ) and four ( $p < 0.001$ ) steps away, but not for targets at one step away ( $p = 0.85$ ). In addition, we found that, for both types of trials, performance decreased significantly from a given target distance to the subsequent target distance but only up to three steps away (feedback trials:  $p_{1-2} < 0.001$ ,  $p_{2-3} = 0.02$ ; no-feedback trials:  $p_{1-2} < 0.001$ ,  $p_{2-3} < 0.001$ ). No significant difference in performance was found between targets that were three and four steps away (feedback trials:  $p_{3-4} = 0.99$ ; no-feedback trials:  $p_{3-4} = 0.99$ ).

We also assessed whether performance improved across the experiment for the different planning horizons. To do so, we implemented a linear mixed-effects model which included a fixed effect for trial number and a random intercept for subjects. The results revealed a significant effect of trial number on the points obtained for both the feedback ( $p < 0.001$ , for all horizons; Figure 3.4B) and no-feedback trials ( $p_{\text{one}} = 0.009$ ;  $p_{\text{two}} = 0.01$ ;  $p_{\text{three}} = 0.04$ ;  $p_{\text{four}} = 0.01$ ; Figure 3.4C) which indicates that the performance improved over the course of the experiment

for all planning horizons. While performances improved with training, it continued to be better for lower planning horizons.

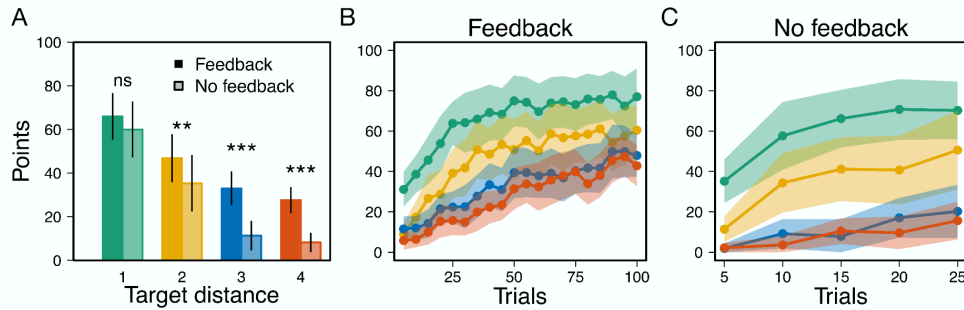


Figure 3.4: Participants performance in Experiment 2 for the different planning horizons. **(A)** Points across the experiment in the feedback and no-feedback trials. **(B)** Points over trials for the feedback trials. **(C)** Points over trials for the no-feedback trials.

Importantly for this experiment, we also examined response times (RT and IKI) for the different planning horizons (Figure 3.5). First, we compared RT for the feedback and no-feedback trials. We carried out a 4X2 repeated measures ANOVA where we found a significant main effect of target distance ( $F(3,72) = 8.35$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.25$ ) and trial type ( $F(1,24) = 15.16$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.38$ ) indicating that RTs were overall higher for the no-feedback trials (Figure 3.5B) and differed by target distance. Post hoc pairwise comparisons revealed that RTs were significantly higher in the no-feedback trials for targets at two ( $p = 0.04$ ), three ( $p = 0.003$ ) and four steps away ( $p = 0.004$ ), but not for targets one step away ( $p = 0.99$ ), which corroborates the increased complexity of multi-step planning without online information.

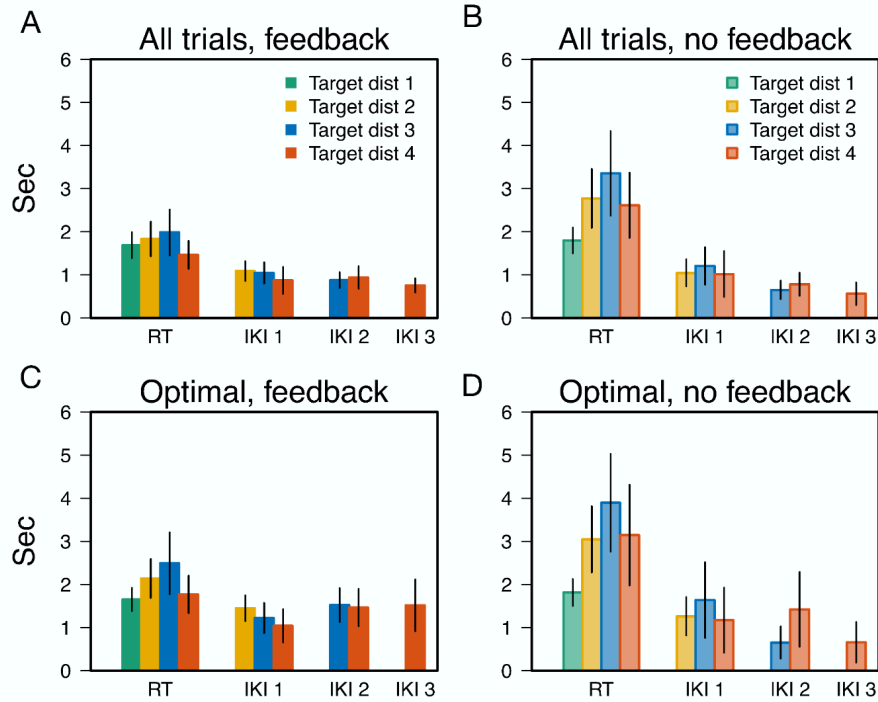


Figure 3.5: Response times in Experiment 2 for the distinct planning horizons. **(A)** RTs and inter-key intervals for all feedback trials. **(B)** Same as in (A) but for no-feedback trials. **(C)** RTs and inter-key intervals for optimal arrivals in feedback trials. **(D)** Same as in (C) but for no-feedback trials.

Additionally, in the trials with no-feedback, where we expected planning to manifest more strongly, we found a significant positive correlation between RT and target distance up to targets three steps away ( $t(73) = 3.2$ ,  $r = 0.35$ ,  $p = 0.001$ ) but not up to targets four steps away ( $t(98) = 1.89$ ,  $r = 0.18$ ,  $p = 0.06$ ). This difference may indicate a potential planning limit in our task, which we investigate further in our modeling analysis (see below). For the feedback trials, we found no significant correlation between RT and target distance, whether up to three steps away ( $t(73) = 1.03$ ,  $r = 0.12$ ,  $p = 0.3$ ) or four steps away ( $t(98) = -0.59$ ,  $r = -0.06$ ,  $p = 0.55$ ), suggesting that they may have been re-planning later on in the sequence and may be the explanation for the lack of difference in performance between distances 3 and 4 steps away.

We also examined response times for optimal trials, where we assumed participants already knew the mapping and mostly relied on planning (Figure 3.5C and 3.5D). Importantly,

our sample size was smaller given that not all subjects managed to arrive at the target optimally, particularly for no-feedback trials. Similar to the analysis for all trials, we found a significant main effect of target distance ( $F(3,59) = 10.49$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.34$ ) and trial type ( $F(1,18) = 8.36$ ,  $p = 0.009$ ,  $\eta_p^2 = 0.31$ ). Notably, post hoc pairwise comparisons revealed that RTs were only marginally higher in the no-feedback trials for the targets at three steps away ( $p = 0.52$ ), but not for the rest of the distances. This suggests that planning was similar between feedback and no-feedback trials when considering optimal performance. We also found that, for the optimal feedback trials, RT increased significantly up to targets three steps away ( $t(73) = 2.52$ ,  $r = 0.29$ ,  $p = 0.01$ ; Figure 3.5C), but not up to targets four steps away ( $t(98) = 0.92$ ,  $r = 0.09$ ,  $p = 0.35$ ). For the optimal no-feedback trials, RT significantly increased up to three ( $t(73) = 4.74$ ,  $r = 0.54$ ,  $p < 0.001$ ) and four steps away ( $t(98) = 3.67$ ,  $r = 0.42$ ,  $p < 0.001$ ; Figure 3.5D).

We did not find any significant difference in the IKI between the trial types or target distances. However, we did find that overall, RTs were higher than the IKI for both feedback ( $t(38.44) = 4.06$ ,  $p < 0.001$ ) and no-feedback trials ( $t(32.28) = 5.42$ ,  $p < 0.001$ ). In addition, the difference between RT and IKI was significantly larger for the no-feedback trials ( $t(33.74) = -4.12$ ,  $p < 0.001$ ), which likely reflected their greater involvement in planning prior to generating the action sequence.

In summary, our behavioral results showed that participants improved performance over the course of the experiment for all planning horizons in both feedback and no-feedback trials. However, their performance decreased for more distant targets and plateaued at targets three steps away. In addition, the fewer points and higher RT in the no-feedback trials for target distances greater than one, highlighted the increased complexity of performing multi-step planning prior to action execution, compared to using online feedback. Lastly, the increase of RT up to targets three steps away in the no-feedback trials, suggested that participants might plan up to this limit. We explore this idea further in our computational modeling analysis.

*Modeling results:* Based on the modeling results of Experiment 1, where the MB-Persev model best described the data, we considered variations of that model which differed in their planning depth (See *Materials and Methods* for details). Specifically, we evaluated models that estimate the value of actions based on simulations ranging from one to four steps ahead. Figure 3.6A shows an example of how the performance of the model can differ depending on the planning depth. If the model considers only one step ahead, it may select suboptimal actions that appear to reduce the distance to the target (green node) based on its heuristic search. However, it can overlook actions that initially seem to move away from the target but ultimately reach it in subsequent steps (gold trajectory). Such options will only be preferred if a deeper search is performed. This model is able to predict the decrease in performance for the distinct planning horizons similar to what we observed in participants (Figure 3.5B). While performance generally decreases with the distance to the target, this effect is attenuated as the depth of the algorithm increases. These simulations were generated using the mean of the best fitting parameters from the Full group in Experiment 1.

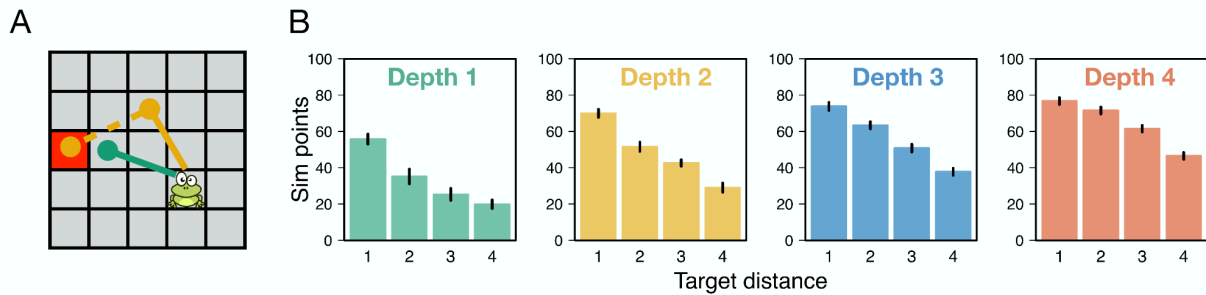


Figure 3.6: Simulations of the MB-Persev model with different depths of search. **(A)** Example of the preferred moves from an algorithm with depth 1 (green trajectory) or depth 2 (gold trajectory). While an algorithm with depth 1 prefers the option that minimizes the immediate chessboard distance to the target, an algorithm with depth 2 prefers an option which is not as close to the target immediately, but leads to it in the next move. **(B)** Simulated points for the different target distances for different depths of search. Overall, performance decreases for more distant targets, but this effect is attenuated as the depth of search increases.

As in Experiment 1, we computed the probability that a random participant from the population was best described by each of our models (Stephan et al., 2009). We found that this probability is 60% for the MB-Persev model with depth 3, followed by 20% with depth 2, and 18% with depth 4. In addition, we found that the best fitting model (MB-Persev depth 3) was able to capture a mean of 40% of the explainable variability in the feedback trials (Shen and Ma, 2016; Grassberger, 2003; Figure 3.7B).

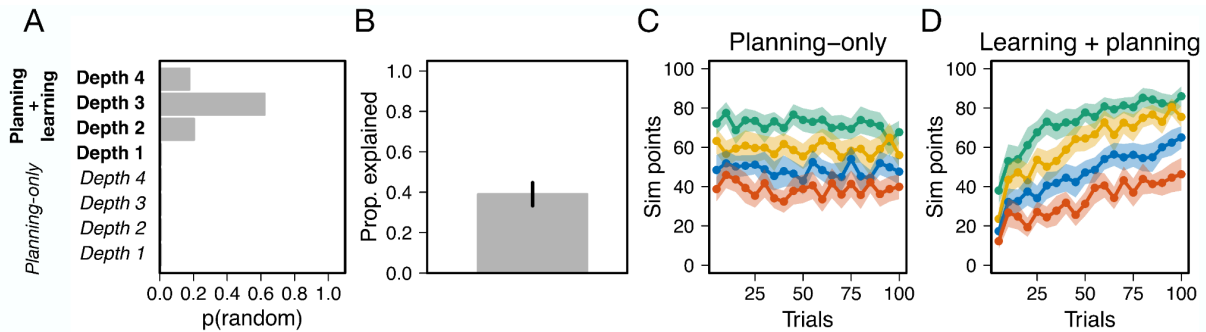


Figure 3.7: Model results of Experiment 2. **(A)** Probability of obtaining each of the MB-Persev models with different depths for a randomly selected subject. **(B)** Proportion of the variability explained by the MB-Persev model with depth 3. **(C)** Simulated points over trials for the Planning-only model (MB) with depth 3 (green = horizon 1; gold = horizon 2; blue = horizon 3; orange = horizon 4). **(D)** Simulated points for the Learning + planning (MB-Persev) with depth 3.

In order to test the importance of mapping-learning in our task, we also considered four “planning-only” models which do not incorporate any learning about state transitions or previous responses into their computations. In such models, the value of actions is entirely the result of the tree search from the planning algorithm with varying depths.

We fitted all models using the responses from the feedback trials and used the parameter estimates to describe the performance in the no-feedback trials. While these planning-only models are able to capture differences in performance for the distinct planning horizon, it assumes that performance remains stable across the experiment (Figure 3.7C) and

does not reflect the learning time course of performance of the participants (Figure 3.4B and 3.4C). In contrast, the MB-Persev depth 3 model, which incorporates learning about the state transitions and about previous actions, is able to qualitatively capture the improvement in performance across the experiment (Figure 3.7D).

We further assessed the performance of the MB-Persev depth 3 model in the no-feedback trials by simulating response probabilities using the observations that participants had in those trials and best-fitting parameters from the feedback trials (Figure 3.8). We only considered responses up to the planning depth for the given trial. For example, if the target was two steps away, we only considered the likelihood of the first two responses. With this, we aimed to exclude responses resulting from guessing, which were likely to happen as the number of responses increased due to the uncertainty about the cursor location. Overall, we found that the model was reasonably accurate at describing responses for targets one step away. However, its accuracy decreases for longer distances, although it performed greater than chance in all cases ( $p < 0.001$ ; Figure 3.8)

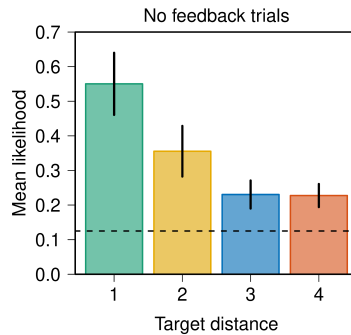


Figure 3.8: Model predictions for the no-feedback trials in Experiment 2. The mean likelihood for the no-feedback trials was obtained using the best-fitting parameters of the model and as many responses as the target distance for the given trial.

In summary, the modeling results of Experiment 2 support our findings in performance and RT, suggesting that participants may plan between 2 to 4 steps ahead in our task, with the

majority of participants performing 3-step planning. In addition, we found that a model that incorporates learning and planning into its computations, best captures the performance improvement over trials as well as the decrease in performance for longer planning horizons.

## **Discussion**

When acquiring novel motor skills, humans face the double challenge of learning what the outcomes of their actions are and 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. Much of motor learning research has focused on sequence learning tasks where the planning component is constrained or limited by the experimenter. On the other hand, planning studies rarely focus on situations where the state transitions require motor learning. In two experimental studies, we aimed to bridge this gap by using a grid navigation task where participants have to learn a novel motor mapping based on the Knight piece from chess, and use it sequentially to arrive at target locations.

In Experiment 1, we showed that performance in this task can benefit from learning the motor mapping first, in the absence of planning. As revealed by our best-fitting model, this training regime can lead to increased learning rates about the state transitions, as well as to a lower influence from habitual responses. The increased performance in a task by breaking it down into its components has been reported previously in the so-called part-whole training (Mané and Donchin, 1989; Park et al., 2010). Practicing a complex component of a perceptual-motor task, such as learning the motor mapping, can yield benefits when the task is later performed as a whole, which can be a useful strategy for acquiring complex motor skills, like in sports. Crucially, our findings suggest that this improvement is partly due to increased learning rates in the state transitions associated with the motor mapping. While we explicitly designed the task so the Mapping group learned the motor mapping first, it raises the interesting

question of how humans break a task into subcomponents on their own (Correa et al., 2023). Our results suggest, however, that without an explicit design, it may be considerably harder to break down a task as indicated by the significantly lower performance of the Full group.

Additionally, in Experiment 2 we showed that participants' performance decreased for longer planning horizons; however, performance significantly improved over the course of the experiment in all of them. This pattern was best explained by a model that incorporates both learning the motor mapping and planning with a depth of three steps ahead. Additional support for this planning depth was found in the significant increase in RT for targets up to three steps away in the no-feedback trials, where planning was expected to manifest more strongly. Although similar results have been reported in “planning-only” tasks (Jensen and Mattar, 2024), our results highlight the critical role of motor learning to capture the improvement of participants across the experiment. In a previous study we have addressed whether prior experience in a planning-only version of our task could benefit the performance when participants learned the mapping and planned, for example, by reducing the cost of planning. However, it did not confer any benefit (Velazquez-Vargas and Taylor, 2023).

Based on previous work on decision-making, we believe the main results from Experiment 1 and 2 could be in part attributed to the distinct working memory loads in the task (Collins and Frank, 2012; Yoo and Collins, 2022). Indeed, in the seminal work from Collins and Frank (2012), the authors show that in an instrumental learning task humans improved their performance at slower rates for larger set sizes of the learned stimuli. A similar effect can be thought to occur in Experiment 1, where the Full group needs to learn an eight-key mapping while potentially attempting to build a decision tree for planning (Huys, 2012). In contrast, the Mapping group only needs to learn the mapping during the first 200 trials, after which they can focus exclusively on planning. Evidence that the Mapping group had a greater involvement in planning when the targets were three steps away, comes from their higher RT compared to the Full group (Figure 3.2B and 3.2C).

In addition, the differences in performance for the distinct planning horizons in Experiment 2 (Figure 3.4B) are remarkably similar to the patterns found by Collins and Frank (2012) for the different set sizes of stimuli. These results suggest that working memory and planning are closely related processes, an idea that has been addressed in recent work (Bottcher et al., 2021; Zhuojun et al. 2023). Indeed, It is conceivable that planning, which is typically thought of as building a decision tree (van Opheusden, 2023; Mattar and Lengyel, 2022) is constrained by working memory capacity. For example, it can limit the size or depth of the tree, or the precision of the reward stored at different levels (Zhuojun et al. 2024). Such constraints can underlie the use of heuristics to reduce the computational demands of planning such as pruning (Huys et al., 2012, 2015) or truncation (Keramati et al, 2016; Krusche et al., 2018) of the decision tree. In our task, our planning model ( $A^*$ ) was constrained in two ways that may reflect the limited cognitive resources in humans. First, we incorporated a limited planning depth. Second, through the incorporation of the chessboard distance as a heuristic function in  $A^*$ , we specified that planning is more likely to occur in the direction of the goal, rather than in any direction such as in uninformed algorithms like Breadth-First Search. These modifications can considerably reduce the working memory load.

While planning depth varies by task, our estimate of three steps ahead lies at the lower end of previously reported planning depths (Huys et al., 2015; van Opheusden, 2023; Arad and Rubinstein, 2012; Snider et al., 2015; Ariani et al., 2021). This can reflect the complexity of planning using the Knight rule, which incorporates non-linear transitions between states and may increase working memory load. Future work can explore how planning differs for mappings with distinct geometries in the state-transitions. For example, by comparing the planning depth while using the Knight mapping and the King mapping. Although both mappings have the same number of actions (eight moves), the latter moves to the adjacent states, which can reduce the complexity of planning and increase the planning depth.

It is important to note that while our best-fitting model qualitatively captures the behavioral patterns in our task, it still falls short in describing the full extent of our data (Figure 3.3B and Figure 3.7B). This could be in part due to cognitive processes that were not specified in our models. However, it could also be attributed to guessing responses, for example, due to attentional lapses. While we did not explore the extent to which participants guessed in our tasks, it could be addressed by extending the model to incorporate a contaminant process, which is common in working memory experiments (Adam et al., 2017; Velázquez-Vargas and Taylor, 2024). Additionally, other planning algorithms could be considered such as Monte Carlo Tree Search, a common planning algorithm in board games (Silver et al., 2016, 2017), which has also been implemented to model how humans plan (Élteto et al., 2023; Krusche et al., 2018).

Finally, while our results provide insights into how motor skills are acquired in discrete state spaces, such as in video games, little is known about how this process occurs in continuous domains. Crucially, a great variety of sequential motor skills, such as playing a musical instrument or sports, involve the control of continuous actions in continuous spaces. These scenarios likely involve the use of internal forward models to simulate the outcomes of candidate actions in combination with internal models about how the world works (McNamee and Wolpert, 2019). Similar to non-motor planning tasks, the dimensionality of the problem suggests the use of constrained algorithms which, in combination with the use of stereotypical or cached solutions (Jax and Rosenbaum, 2007), can make planning tractable.

Overall, the present studies show the critical role of integrating motor learning into planning tasks, which is a crucial step to understanding how humans acquire a wide and complex set of motor skills throughout life.

## Supplementary Information

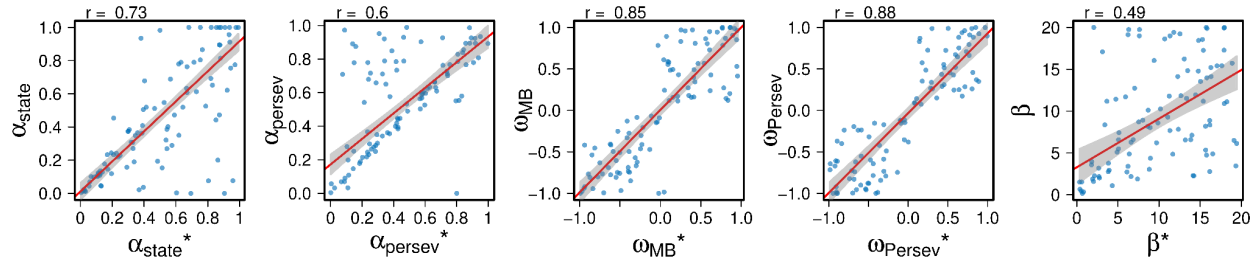


Figure S3.1: Parameter recovery of the best fitting model (MB-Persev) using a depth of three steps ahead.

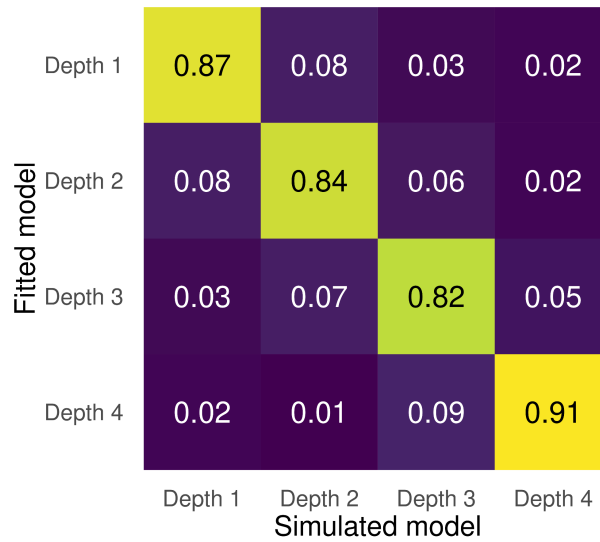


Figure S3.2: Confusion matrix with model recovery results. Each row and column shows the MB-Persev model with different depths of search. Numbers inside the cells represent the proportion of times that the model in the Y axis best recovered the data generated by the model on the X axis according to BIC.

## Chapter 4

### Introduction

Chapters 2 and 3 provided insights into key processes involved in acquiring a new motor skill: learning a novel mapping and planning with it. In this chapter, I address a third challenge, commonly encountered in various learning scenarios: how to prevent interference when multiple tasks (i.e., mappings) are being learned simultaneously? Furthermore, how do contextual cues help in preventing such interference?

There has been considerable interest in determining how contextual cues allow the consolidation and retrieval of multiple visuomotor memories (Howard et al., 2013; Heald et al., 2018; 2023a). While it has been shown that arbitrary external contextual-cues – such as colors, sounds or shapes – are effective in separating the task contingencies in a variety of domains such as in classical conditioning (Gershman, 2017), episodic memory (Pu et al., 2022) and value-based decision making (Bornstein & Norman, 2017), the same cues cannot prevent catastrophic interference in visuomotor adaptation tasks (Howard et al., 2013) unless they are presented in close temporal proximity (Avraham et al., 2022).

In contrast with these laboratory-based findings, we appear capable of storing multiple mappings when using a variety of digital devices despite having similar movements (e.g., video games use the same controller to play a racing game and a first-person shooter). One interpretation of the above, is that a great proportion of these mappings may not be the result of visuomotor adaptation but instead of *de novo* skill learning. When a skill is acquired *de novo*, new control policies are created, rather than recalibrating existing ones (Krakauer et al., 2019).

In a typical *de novo* task, participants are required to learn arbitrary, and usually non-intuitive associations between their movements and the outcomes to achieve the task goals (Wilterson & Taylor, 2021; Mosier et al., 2005; Velazquez-Vargas & Taylor, 2023). The grid-navigation task developed in chapter 2 presents a model paradigm to study how multiple

mappings learned *de novo* can be consolidated while preventing interference. Crucially, *de novo* tasks are known to involve brain regions associated with declarative processes such as the hippocampus (Wise & Murray, 2000) and the prefrontal cortex (Fermin et al., 2016), also crucial for context-dependent learning (Heald et al., 2023b). This important distinction could allow *de novo* skill learning to be sensitive to contextual cues that are efficient in domains such as episodic memory, unlike visuomotor adaptation tasks which are known to involve a significant cerebellar-dependent component (Taylor et al., 2010).

Meta-learning (Wang et al., 2018) can be a unifying computational framework that formalizes *de novo* skill learning. In meta-learning, an agent (human or artificial) has to learn a distribution of different but related tasks. For *de novo* skill learning, the different tasks can use the same motor repertoire but may map each motor movement to a different outcome. Within this framework, a meta-learning agent can either learn a single abstract representation across all tasks that accommodates all motor mappings or learn separate representations for all the unique motor mappings seen across tasks, where each representation can be bound to a particular external contextual cue (Musslick & Cohen, 2021). Meta-learning models are becoming increasingly more relevant in cognitive modeling of behavioral phenomena due to their ability to implement Bayes-optimal learning algorithms on tasks for which Bayesian inference is intractable (Binz et al., 2023).

In the present study, we designed a grid navigation task to dissociate these two hypotheses. One group of participants (the context group) performed a grid navigation task (Fermin et al., 2010; 2016; Velazquez-Vargas et al., 2023) where they moved a cursor from start to target locations in two “worlds” randomly interleaved over trials (Figure 1). Each grid-world was associated with a unique key-mapping and had distinctive contextual cues. We compared performance on this group with another group of participants (no-context group) that experienced the trial-changes in key-mappings but not in the external contextual cues –i.e., grid worlds.

To implement cognitive models for learning to navigate with different visuomotor mappings, we trained two recurrent-based meta-learning agents using the architecture from Wang et al. 2018, which is an LSTM agent trained with reinforcement learning using an Actor-Critic framework (Mnih et al., 2016). The first model (context LSTM) incorporated external contextual information into its input while the second one did not (no-context LSTM). We tested their performance in the task and how well they predicted human participants' data. In addition, we examined how their internal representations gave rise to different strategies to learn multiple key-mappings, giving us potential insights and hypotheses for the neural representations of the participants.

We found that humans and the meta-learners performed better in the task when provided with contextual information. We also found that, in both conditions, there is individual variability as to whether an individual acts more like the context LSTM or the no context LSTM. Additionally, we found that when no contextual information is provided, the internal representations of the LSTM agent are highly correlated in space and time while using the different key-mappings. However, these representations are less correlated when the LSTM agent has the external context input. This suggests that the former is learning separate representations for different motor mappings whereas the latter is binding different motor mappings to the same representation. Finally, we showed that the capacity to learn multiple key-mappings is dependent not only on the presence of contextual cues, but also on the complexity (number of hidden LSTM units) of the internal representations of the model. Our model-based neural analyses and human behavior results give us potential insight on how the brain can effortlessly learn multiple visuomotor mappings.

## **Materials and Methods**

*Participants:* Thirty two participants from Princeton University (13 males and 19 females, mean age = 21.7, sd = 3.8) were recruited through the psychology subject pool. The study was

approved by the Institutional Review Board and all subjects provided informed consent prior to doing the experiment.

*Task:* The experimental task was programmed using HTML/CSS/Javascript and hosted on Google Fire- base. Visual stimuli were displayed on a 60 Hz Dell monitor using a Dell OptiPlex 7050'a machine (Dell, Round Rock, Texas) running Windows 10 (Microsoft Co., Redmond, Washington). Participants' responses were recorded using a standard desktop keyboard. Before the start of the experi- ment, the instructions were displayed on the screen.

The goal of the task was to move a cursor from start to target locations in a 9X9 grid environment using the least possible number of moves (Fermin et al., 2010; 2016; Velazquez-Vargas et al., 2023; Bera et al., 2021). To do so, participants used three keyboard keys (J, F and K) that moved the cursor to arbitrary and unknown adjacent locations (key-mapping): left-up, left-down and right (Figure 4.1). The task consisted of 360 trials where the start-target locations were randomly sampled out of eight possible pairs. For each pair, the target was placed seven moves away from the starting location of the cursor. If participants arrived at the target using the minimum number of moves (optimal arrival), they would observe a happy emoji face at the target location and hear a pleasant sound. If they arrived at the target but not using the minimum number of moves (arrival), they would observe a neutral face and hear a neutral sound. If they did not arrive at the target in less than 10 s, the trial was terminated and participants would observe a sad face and hear an unpleasant sound.

Importantly for this study, participants performed the tasks using two key-mappings, which corresponded to different arrangements of the same moving directions (see Figure 4.1), and which were randomly interleaved across trials. Participants were not informed that there would be more than one key-mapping to perform the task. In order to test the effectiveness of external contextual cues in separating the learning of the visuomotor mappings, we split participants into two experimental groups.

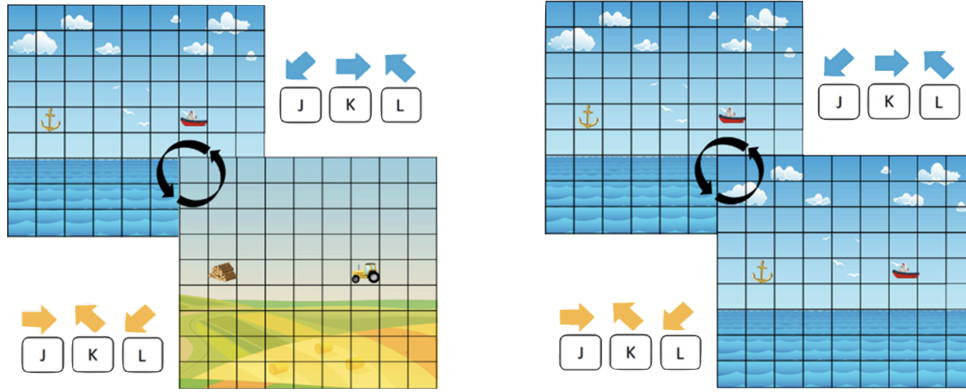


Figure 4.1: Experimental task Subject performs a grid navigation using different key mappings randomly interleaved over trials. In the context group, the key-mappings were deterministically signaled by a unique grid world, whereas in the no-context group, participants used both mappings in the same world.

**Context Group:** In this group ( $n = 16$ ) participants performed the grid navigation task in two different “worlds”, either in the ocean or in a farm. Each world consisted of unique visual (background, cursor shape) and auditory (ambience song and cursor sound when moving) information. Most importantly, each world was deterministically associated with one of the key-mappings.

**No-context Group:** Participants in this group ( $n=16$ ) performed the grid navigation task in a single world while still experiencing the changes in the key-mappings. At the beginning of the experiment, one of the worlds would be selected and remained throughout the whole task.

**Models:** Following previous work on meta reinforcement learning (Wang et al., 2018; Kumar et al., 2020), we used a Long short-term memory (LSTM) network trained using Advantage Actor Critic (A2C; Mnih et al. 2016). The recurrent structure of this agent, joint with a model-free learning mechanism, makes it suitable for sequential tasks like ours, where trial-and-error, reward-based processes are crucial to achieve the goals. The agents were implemented using Stable baselines 3 package (Raffin et al., 2019) with the Proximal Policy Optimization algorithm (PPO). We set constraints of the hyperparameters of PPO so that the

algorithm becomes equivalent to A2C. We used Optuna (Akiba et al., 2019) to tune the following hyperparameters: learning rate, discount factor gamma, GAE lambda, number of steps, batch size, entropy coefficient, the value function coefficient and the number of LSTM hidden units. After optimization, we trained the agents for  $2 \times 10^5$  time steps, where they reached asymptotic performance. The reward structure was defined as follows: every move that did not lead to a terminal state (arrival to the goal) received +1, except when colliding with the grid walls, in which case the agent received a penalty of -100. If the agent reached the goal, it received a reward of +10.

We trained two types of agents that differ chiefly in the input provided to the LSTM network. The first one (context LSTM) received its current location and target locations on the grid as well as a context vector indicating which of the worlds was currently being observed. This agent faced the same situation as participants in the context group where the key-mapping changes were linked to changes in the contextual cues. On the other hand, the second agent (no context LSTM) only received its current location and the target location as input, mirroring the design of the no-context group.

## Results

*Humans and models' behavior:* As a metric of human performance in the context and no-context groups, we provide the proportion of times they arrive at the target optimally – i.e., using the minimum number of key presses– across trials and in the entire task. Based on the optimal arrivals to the target, we observed that participants in both the context and no-context groups improved their performance over trials (Figure 4.2). However, the context group performed significantly better overall ( $p = 0.03$ ). For the context LSTM and no-context LSTM agents, we provide the average reward per episode at different evaluation points across training, as well as at the end of training. The context LSTM and the no-context LSTM agents

showed the same pattern as participants, where both models improved over the timesteps but the former performing overall better in the task ( $p < 0.05$ ).

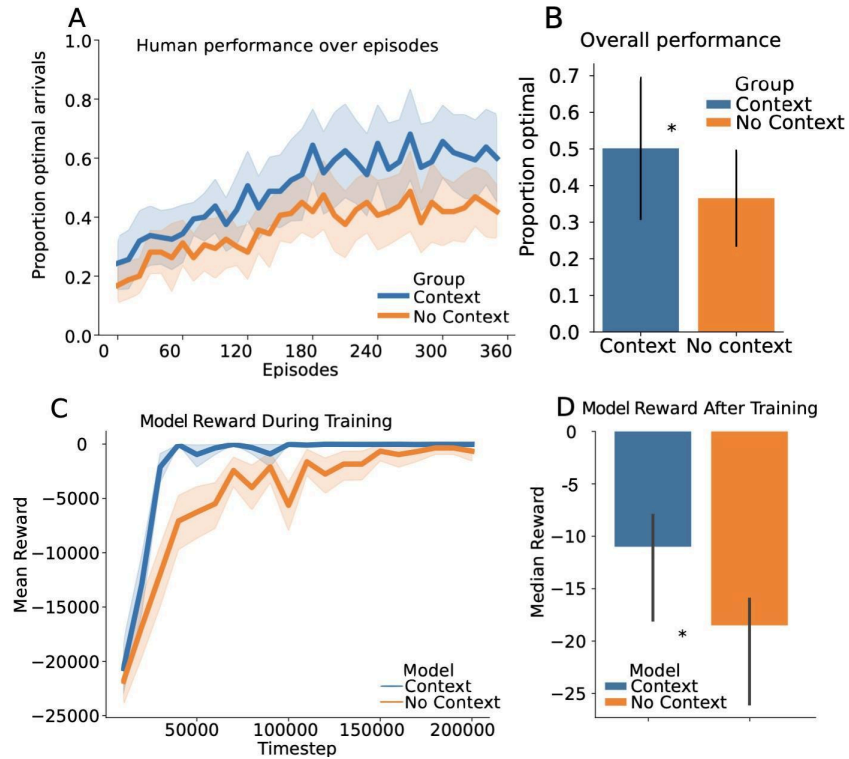


Figure 4.2: **Performance of Humans and Meta-RL Agents** (A). Mean performance of humans over the episodes they completed, measured by proportion of optimal arrivals. Shading represents 95% confidence intervals. Humans that were given context cues learned the task better. (B). Overall performance differences across different experimental groups in humans. Those in the context group did significantly better. (C). Mean reward over time during agent learning for both context and no- context agents. (D). Median reward after training. Agents that were given an external contextual cue for the mapping as additional input did significantly better.

To identify the similarity between subjects' behavior and the context LSTM and no-context LSTM agents, we calculated the likelihood of subjects' responses according to the fully trained agents. That is, we performed a forward pass on the agents and obtained the likelihood that they would take the same action as the human. The likelihood for each action was averaged across all timesteps for each subject to produce one value indicating the average likelihood of a subject's response under the model. This analysis was done for both the context

and the no-context LSTM agents.

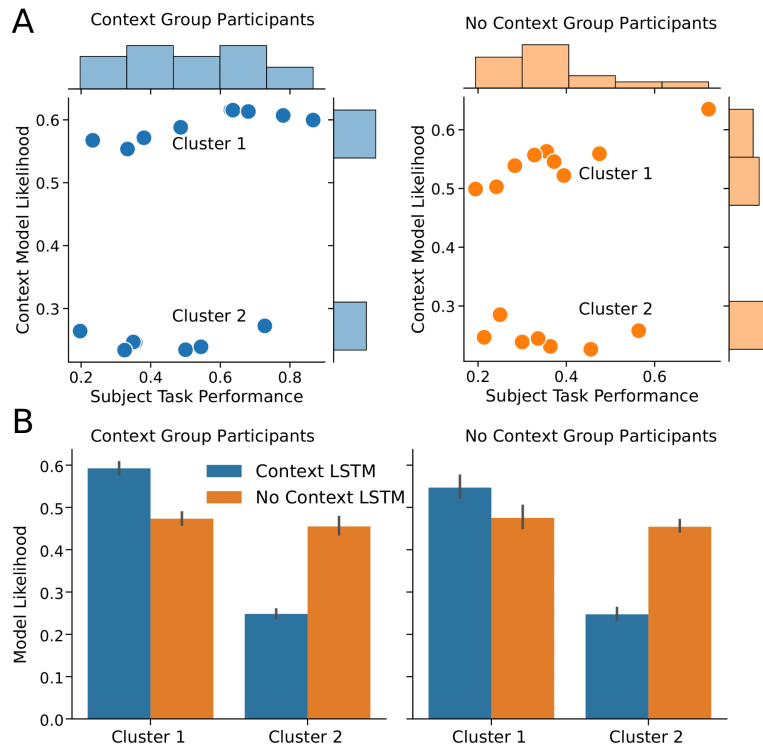


Figure 4.3: In both experimental groups, humans are split between behaving like the context vs no context LSTM (A). Joint scatterplot and histograms for likelihood of subjects' actions under the context LSTM model vs. performance of subjects on the task. In both experimental groups (whether participants received external context input or not), there are two clusters of participants — those whose actions are explained well by the context LSTM model and those whose actions are not. (B). If we examine the mean likelihood under both types of models (context vs no context LSTMs), we see that participants whose actions aren't as well explained by the context LSTM model have significantly higher likelihood under the no context LSTM model.

Figure 4.3A shows the average response likelihood for both experimental conditions under the context LSTM agent, where two clear clusters were formed. In Cluster 1, the context-LSTM agent was considerably better in predicting participants' responses than in Cluster 2, where it performed around chance level. Following this analysis, we found that participants that were poorly predicted by the context LSTM (Cluster 2) agent were instead significantly better predicted by the no-context LSTM ( $p < 0.05$ ; Figure 4.3B). Likewise, for participants in Cluster 1, the context LSTM agent better predicted participants' responses

compared to the no-context LSTM agent.

These results indicate that regardless of the experimental group, some participants were better predicted by the context model, suggesting that even in the absence of contextual cues from the experiment, participants could have relied on alternative cues to separate the representation under the different key-mappings. A point we will address in the Discussion section.

*Model representation similarity analysis:* We employed multiple analyses (Figure 4.4) to examine differences in representations between the cue and no-cue agent. Specifically, we first investigated how each network represents the same trial and environment under different motor contexts. To do this, we performed a Representational Similarity Analysis (RSA; Kriegeskorte et al. 2008) on both a spatial and temporal scale. For the spatial RSA (Figure 4.4A), we evaluated the agents in  $10^4$  episodes and computed the average hidden state (LSTM units) at each location in the 9x9 grid, separating the hidden states according to the key-mapping being used. For every grid state, we obtained the correlation coefficient between the averaged hidden states when using key-mapping A and the averaged hidden states when using key-mapping B. Lower correlations would indicate that agents represent the grid space between different action mappings more distinctively. In the context LSTM agent, the correlations were low across most grid states, with many locations on the grid showing no correlation, suggesting that spatial location is represented differently when using mapping A and mapping B. In contrast, for the no-context LSTM agent, the RSA generally depicts grid states with correlations that are significantly higher ( $p < 0.001$ ), when the different key-mappings are being used compared to the context LSTM agent. The difference between the two suggests that the context LSTM agent has a distinct representation for grid states that depends on its current context, while the no-context LSTM model does not.

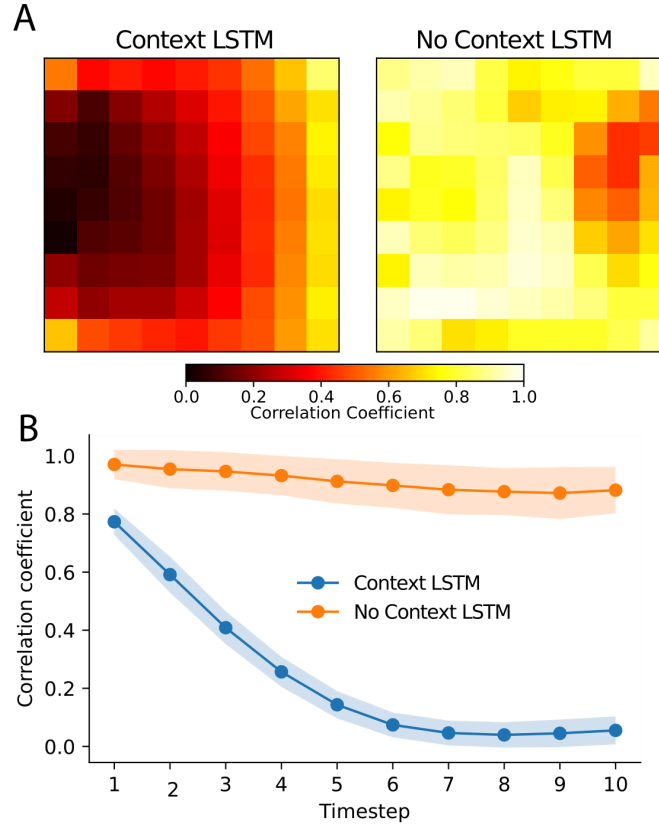


Figure 4.4: Context agents have lower representational similarity under the different key-mappings over space and time (A). Spatial RSA analysis. We correlated LSTM hidden state representations of the same episodes under the different key-mappings and showed the mean correlation across different spatial locations in the 9 9 grid. For context agents, this correlation is much lower, presumably because the representations of different key-mappings are more separated. (B). Temporal RSA analysis. We show how the correlation changes over time by plotting, for each time-point, the mean correlation of LSTM hidden state representations under the different key-mappings. For context agents, this correlation goes down over time whereas for no-context agents, there is less change overtime. This suggests context agents exhibit more dissimilar representational similarity under different key-mappings overtime.

The temporal RSA (Figure 4.4B) provides complementary results showing that context is represented differently between models not just in space, but also in time. Specifically, we explored the similarity of hidden states of the agents as they progressed towards the target. To do so, we evaluated the agents in  $10^4$  episodes and for every timestep we correlated the averaged hidden states when using key-mapping A with the averaged hidden states when using key-mapping B. In the no context LSTM agent, the representations on the initial timestep are

highly correlated. As the agent steps towards the goal, representations become slightly less correlated, but for the most part remain highly correlated across the episode. In the context LSTM agent, however, representations under the different key- mappings drastically change with each step in the episode. Representations begin similarly on the first time step (although less so compared to the no-context LSTM agent), but quickly drop to no correlation at around the 6th step in the episode. This suggests that the agent representing context increasingly leverages its context representation as it nears the goal state. Taken together, these results show two different representational strategies, one where the agent attempts to learn distinct representations for each context, and another where the agent learns the task without an explicit representation of context.

*Model representation capacity analysis:* Given the capacity limitations in memory, computational resources and time, we would expect that humans exhibit constraints in the number of visuomotor mappings they can learn. To explore this idea, we investigated the agents' capacity to learn multiple key-mappings by varying the complexity of their internal structure, i.e. number of hidden units in the LSTM. In order to do so we exposed both the context LSTM and no-context LSTM agents to up to ten different key-mappings and used the following number of hidden units: 2, 5, 10, 20, 40, 80, 100, 120, 256 or 512. Each agent was trained for 200k timesteps followed by 100 evaluation episodes from which we report the averaged (normalized) reward.

Figure 4.5 displays performance of both models as the number of contexts and representational capacity increases. Across both models, we see better performance with more representational capacity. Between models, the model that explicitly receives contextual cues performs better as contexts increase than does the model that does not receive a cue. This is perhaps due to the way in which context is represented between both models. In the context model, each context is distinctly represented within the hidden layer, allowing for a unique task

representation for each context. In contrast, the no context model does not distinctly represent each context. This supports work showing that unique task representations are necessary as the environment becomes more complex (i.e., when the number of contexts increase) and adds that explicit cues aid in both the speed at which the task is learned and how it is represented.

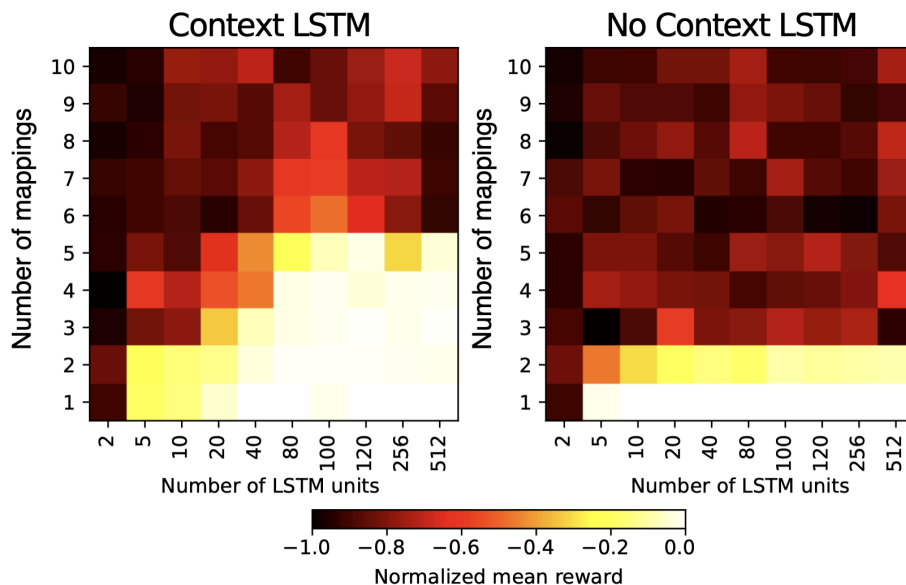


Figure 4.5: For context agents, scaling the LSTM capacity enables learning more mappings For both context and non-context agents, we varied the number of hidden units and exposed them to different numbers of mappings during training. For context agents, increasing the number of LSTM units allows for learning of more mappings until about 5. For no-context agents, varying the number of LSTM units does not have as strong an effect.

## Discussion

Humans possess a remarkable capacity to learn complex, and often arbitrary, visuomotor mappings. For example, when learning to play video games humans are able link the button presses from the controller with arbitrary actions in the game. What is more, not only are they capable of learning such arbitrary mappings and use them to achieve goals, but they can build an entire repertoire of them. It is not uncommon to witness players seamlessly transitioning between mappings for distinct gaming genres, be it from a first-person shooter to a

racing or football game.

In this chapter, we have shown that the acquisition of distinct visuomotor associations can be aided by external contextual cues. In particular, participants that are provided with contextual information that differentially cues the key-mappings, performed significantly better than participants without it. These results are supported by previous findings outside the motor domain where external contextual cues allow subjects to discriminate between different contingencies of the task (Gershman, 2017; Bornstein & Norman, 2017). Similar cues, however, do not prevent catastrophic interference in visuomotor adaptation (Howard et al., 2013), except under specific circumstances (Avraham et al., 2022). We believe this effect arises due to the fundamental distinction between adaptation and *de novo* learning, where the former recalibrates an existing control policy and the second one builds one from scratch. This can generate qualitative differences between the contextual cues the systems supporting each process are sensitive to. On the one hand, cerebellar dependent recalibration processes could be more easily influenced by contexts related to the body states or the world kinematics Howard et al. (2013). However, *de novo* learning, particularly in the domain of spatial navigation likely involving the hippocampus and the prefrontal cortex (Fermin et al., 2016), could be more sensitive to visuospatial contextual cues.

It is important to emphasize that while the context group outperformed the no-context group, the latter still improved its performance over trials, arriving optimally a greater proportion of times toward the end of the experiment ( $p < 0.05$ ). This improvement occurred in the absence of external contextual cues, which could aid in differentiating the key-mappings. We believe this result could be attributed to participants relying on a different set of cues: movement-related ones. In particular, given the specific movement directions that we used for the key-mappings (left-up, left-down and back), it is possible to arrive at the target locations with multiple and equivalently optimal trajectories which begin with a different action. For example,

for a target that is right above the cursor, pressing the key that moves right, or left-up, would leave the cursor one move away from the target. Although this move makes no difference in terms of performance, it does provide information about the ongoing mapping, given that no key moves the cursor to the same direction. Therefore, participants could have used the first move of the cursor to sample the ongoing key-mapping and adapt the subsequent moves accordingly. In effect, this strategy would have provided participants with contextual information similar to the context group and would have been able to learn the two mappings separately. This is consistent with the fact that some participants in the no-context group had a higher likelihood under the context LSTM than the no context LSTM (Figure 4.3A). Alternatively, participants could have learned a single of the two key-mappings, or an averaged version of them, which over time would have allowed them to improve in the task, although not to the same extent as the context group. Participants relying on the latter strategies would be learning, in effect, a single key-mapping. Further analysis would be needed to test these hypotheses.

In order to have a behavioral model for the presence or absence of contextual cues while acquiring multiple mappings, while also providing insight into the neural representations, we leveraged recent developments on meta-reinforcement learning (Wang et al., 2018; Huang et al., 2022; Kumar et al., 2020). Mirroring human performance, we found that an LSTM agent that explicitly receives information about its context outperforms an LSTM agent that does not. Moreover, we found that participants that were poorly described by the context LSTM, either in the context or no-context groups, were instead best described by an LSTM model with no contextual information. Through implementing these models, we were able to find that within both experimental groups, people vary in either producing behavior consistent with a model learning separate vs shared representations across different mappings (Figure 4.3).

Based on the spatial and temporal RSA, we would expect to see differences within neural representations of participants relying on contextual information (either from the

experiment or potentially from their own actions) vs. individuals that don't. However, due to the fact that the no context LSTM explained some participants' behavior better than the context LSTM (Figure 4.3), it is conceivable that some participants held a single representation in the presence of contextual cues, while others held different representations in the absence of them. This may beg the question — why would a participant choose to hold a single representation even when given contextual cues? The fact that the number of mappings learned scales with the capacity of the model (Figure 4.5) suggests that learning these separate representations uses more cognitive resources. Subjects choosing to not learn separate representations may be behaving rationally according to their cognitive resources (Lieder & Griffiths, 2020). From our capacity analysis of the networks in Figure 4.5, we predict that participants who don't learn separate representations may take longer to learn the same number of mappings than participants who take advantage of the contextual cues to learn separate representations. Confirming these predictions with brain imaging, and generally studying how participants share vs. separate neural representations of visuomotor mappings to manage time and computational resources, will be a rich line of future work.

## Chapter 5

### General discussion

When acquiring a novel motor skill we often face several challenges. For example, imagine that you are learning to use your first video game console. In order to become skillful in the games you will play, you will first need to figure out what the outcomes of pressing the buttons are, which are likely to be highly arbitrary in some games. The association between pressing the buttons and their outcomes constitutes a visuomotor mapping which will set the basis of your future skill. Throughout your learning process, you can first familiarize yourself with the functions of the buttons. Eventually, you may begin to memorize the sequences of button presses that were successful in previous game states, and form cost-efficient stimulus-response associations (Logan, 1988).

In Chapter 2, I examined these two types of mappings—function-based and stimulus-response associations—using a grid navigation task originally developed by Fermin et al. (2010). Unlike typical sequence learning experiments, Fermin et al. (2010) task allows the study of internally-generated sequences (Bera et al., 2021) with a novel mapping, serving as a model for investigating visuomotor skills in discrete action spaces, such as those found in various digital devices, including video games. The general finding from this chapter was that training variability induced either by the number of start-target pairs or the stochasticity of the mapping, led to successful generalization and likely reflected the use of function-based mappings over stimulus-response associations. I formalized these results into reinforcement learning models which shed light into the mechanistic nature of the mappings. In particular, a trial and error process that constructs state-action values (Sutton and Barto, 1998; Niv and Schoenbaum, 2008; Miller et al., 2019; Schultz et al., 1997) was used to represent mappings as in stimulus-response associations, whereas a processes entailing simulation from a model of

the environment (Glächer et al., 2010; Daw et al., 2011; Tolman, 1948; Behrens, 2018) represented the mappings based on the function of actions.

While training variability usually leads to better generalization in several domains (Raviv et al., 2022), this outcome is not consistently observed in standard motor learning experiments (Berniker et al., 2014). This is likely because a great number of studies in motor learning focus on the implicit component of it, which is known to proceed rather locally (Mattar and Ostry, 2010; Berniker et al., 2013) and through cerebellar-dependent mechanisms (Taylor et al., 2010). In contrast, the grid-navigation task involves processes, such as spatial navigation and planning, which rely on brain structures typically regarded as declarative or deliberative, such as the hippocampus (O'Keefe and Nadel, 1978) and the prefrontal cortex (Unterrainer and Owen, 2006). This distinction may have provided participants with the mechanisms to develop successful generalization when exposed to training variability, and placing them in the terrain of more cognitive tasks (Raviv et al., 2022). The grid navigation task, therefore, serves as a promising testbed for exploring the relationship between motor learning and non-motor processes.

Internally-generated motor sequences, i.e. when no instruction is provided about the sequence to use to reach a goal, seem intuitively common in many human activities. However, they have remained widely understudied. A key component of these sequences is that they require planning. Common planning tasks in humans include board games (van Opheusden et al., 2021), graph navigation (Callaway et al., 2022), grid navigation (non-motor; Jensen and Mattar, 2024) or instrumental multi-step tasks (Daw et al., 2011). Importantly, to my knowledge none of these existing tasks involves motor-learning even if some of them like the two-step task by Daw et al.(2011) do require learning about the state transitions. This contrasts with how motor skills are acquired, which require sequential decision making but also the learning of a novel motor mapping. For example, in a fighting video-game, a player not only needs to figure out what the buttons do (e.g., kick, jump, grab, duck, etc), but also how to concatenate them to

generate better attacks or “combos”. Chapter 3 focuses on this question: how does motor learning interact with planning?

My findings from this chapter indicated that starting with the motor learning component of the task, specifically learning a mapping based on the Knight chess piece, followed by its use in sequential decisions leads to better performance in the task. This training regime outperformed an alternative approach where participants learned the motor mapping while simultaneously attempting to reach targets requiring sequential decisions. These results are likely to occur because the working memory load is lower when focusing only on learning the mapping, rather than when simultaneously attempting to arrive at the distant target, which presumably involves building a decision tree (Huys, 2012). Indeed, previous work has shown that learning proceeds more slowly for greater working memory loads (Collins and Frank, 2012; Oberauer et al., 2018). The cognitive constraints from participants in my task were also observed in their decreased performance for more distant targets and the increased RTs. I found that a model with a depth-limited planning component was able to capture the decline in performance. However, it also required a motor learning component to capture the general improvement in the task for all planning horizons. These results, therefore, highlight the importance of both motor learning and planning for a sequential decision making task.

The capacity constraints found in the studies from Chapter 3 have an interesting similarity to the ones I found in a visuomotor adaptation task where instead of planning, participants recalled visuomotor solutions (Velázquez-Vargas and Taylor, 2024). In particular, I found that humans can store between 2-5 items to solve an adaptation task, similar to other working memory studies (Oberauer et al., 2018). While standard working memory tasks differ from the studies reported in this dissertation, the capacity limitations have a strong similarity with the limited depth of search in planning found here and in previous studies (Huys et al., 2015; van Opheusden, 2023; Arad and Rubinstein, 2012; Snider et al., 2015; Ariani et al., 2021). This suggests that planning and working memory are closely related. It is conceivable that working

memory can constrain planning in multiple ways, for example by limiting the depth, truncating unpromising branches or limiting the precision of the reward information at each state (Zhuojun et al. 2023). I believe the study of how working memory limits planning processes is a fruitful line of future research.

In addition, as addressed in Chapter 3, it remains an open question how planning in discrete spaces like the one in my task—and the one in most non-motor planning tasks—relates to planning in continuous domains such as in reaching movements. While tree search algorithms are central for discrete states, Optimal Feedback Control algorithms have been dominant in describing movement planning and execution (McNamee and Wolpert, 2019). However, most complex motor skills are likely a mixture of planning in abstract (discrete) spaces with the planning of continuous movements. It remains an open question how these two types of planning relate to each other and whether they interact.

In the final project of this dissertation (Chapter 4), I explored the challenge of learning multiple mappings simultaneously and examined whether external contextual cues can prevent interference between these mappings. For instance, when learning to play two different open-world games using the same controller, humans might use sensory cues—such as characters, backgrounds, or auditory information—to separate the internal representations of the mappings. In particular, I proposed that this process can occur using meta reinforcement learning. In meta reinforcement learning, humans and artificial agents learn to abstract common features from separate but related tasks, and transfer them to novel tasks (Wang et al., 2018; Kumar et al., 2020). I tested this framework in a dual-learning task of grid navigation where I found that both humans and meta reinforcement learning agents performed better with contextual cues. In addition, through model-based simulations of the artificial agent, I found that contextual cues allows the separation of the internal representation of the mappings in space and time, and enable the learning of significantly more mappings (around five) than in the absence of these cues, where presumably the contexts need to be inferred.

While it can seem intuitive that contextual cues can be used to separate the representations of different mappings, this is not always the case in motor learning experiments — particularly in tasks dominated by implicit adaptation processes (Howard et al., 2013). Indeed, when external contextual cues such as colors are used to adapt to opposing force fields, participants often exhibit catastrophic interference (Howard et al., 2013). Similar to the findings on generalization in Chapter 2, the contrast between the limited effectiveness of external contextual cues in implicit adaptation tasks and their advantage in the grid navigation experiment, may arise because the latter engages mechanisms typically associated with declarative tasks.

Meta reinforcement learning can provide valuable insights into how knowledge across multiple motor skills is transferred to novel ones. For example, a player that is good at shooting games may be more likely to be good at a new one, even if the mapping of the controller changes, compared to a player that mostly plays open-world games. While I did not test a similar scenario of transfer in my tasks, future work can explore whether participants trained with a particular set of mappings have a learning advantage in learning a new one from the same set compared to participants trained in a different set or with no prior practice at all.

Further assessment of these meta reinforcement learning models is necessary to see if they provide an accurate account of how humans learn their repertoire of motor skills. For example, it remains to be evaluated how they complement previous models of contextual inference that rely on generative models and Bayesian inference (Heald et al., 2021).

Lastly, it has been debated whether sequence learning tasks, including grid navigation, truly measure motor learning or if they are primarily visuospatial learning tasks (Krakauer et al., 2019). While the visuospatial and motor components of the grid navigation task by Fermin et al. (2010) have not been dissociated, this question has been addressed in simpler sequence learning experiments. In particular, Verwey and Wright (2004) asked participants to learn sequences of key presses with one hand and then tested them with novel or the familiar

sequences either with the same or the other hand. The authors found that familiar sequences were performed faster with the hand they had used during practice than with the new hand, suggesting an effector-dependent effect. If learning had been purely visuospatial, the familiar sequences would have been executed equally good with both hands. Additionally, an effect-independent effect was observed, where participants performed familiar sequences faster than unfamiliar ones when using the unpracticed hand, indicating the presence of visuospatial learning. Therefore, it is likely that for the grid navigation task there exists a contribution of both visuospatial and motor components.

A similar point can be raised about whether the mappings learned in the task are truly *de novo* for the participants. While it is likely that participants first experience the association between the particular keys and their outcomes during the experiment, they could have had preexisting knowledge that facilitated the learning process. For example, participants can have inductive biases to assume that the cursor will move to contiguous locations rather than distant ones when a key is pressed, or that different keys do not move the cursor to the same location. In Chapter 3, some participants reported being familiar with the mapping of the Knight, which could have facilitated the planning process even if they had never associated it with the keys of their keyboard. Future work can study these biases or preexisting knowledge and whether it makes some mappings easier to be learned than others. For example, mappings that are symmetric (e.g., the Knight mapping) may be learned faster than asymmetric ones (e.g., a mapping that moves the cursor to eight random locations).

Overall, the studies presented in this dissertation aimed to improve our understanding about how humans acquire new motor skills, with an emphasis on discrete action spaces and three key processes involved in their development: mapping-learning, planning and context learning. While these processes have been studied separately across different domains, my approach was to explore them within a single model paradigm—grid navigation. This framework

revealed intriguing findings that contrast with established results from motor adaptation experiments.

## References

Abrahamse, E. L., Ruitenberg, M. F., De Kleine, E., & Verwey, W. B. (2013). Control of automated behavior: insights from the discrete sequence production task. *Frontiers in Human Neuroscience*, 7, 82.

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. In *Advances in Neural Information Processing Systems* (Vol. 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.

Adam, K. C., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, 97, 79-97.

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, 34(2), 77-98.

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).

- Arad, A., & Rubinstein, A. (2012). The 11–20 money request game: A level-k reasoning study. *American Economic Review*, 102(7), 3561-3573.
- Ariani, G., Kordjazi, N., Pruszynski, J. A., & Diedrichsen, J. (2021). The planning horizon for movement sequences. *ENeuro*, 8(2).
- Avraham, G., Taylor, J. A., Breska, A., Ivry, R. B., & McDougle, S. D. (2022). Contextual effects in sensorimotor adaptation adhere to associative learning rules. *Elife*, 11, e75801.
- Balsters, J. H., & Ramnani, N. (2011). Cerebellar plasticity and the automation of first-order rules. *Journal of Neuroscience*, 31(6), 2305-2312.
- Bapi, R. S., Doya, K., & Harner, A. M. (2000). Evidence for effector independent and dependent representations and their differential time course of acquisition during motor sequence learning. *Experimental Brain Research*, 132, 149-162.
- Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, 100(2), 490-509.
- Bera, K., Shukla, A., & Bapi, R. S. (2021). Cognitive and Motor Learning in Internally-Guided Motor Skills. *Frontiers in Psychology*, 12, 604323.
- Berniker, M., Franklin, D. W., Flanagan, J. R., Wolpert, D. M., & Kording, K. (2014). Motor learning of novel dynamics is not represented in a single global coordinate system: Evaluation of mixed coordinate representations and local learning. *Journal of Neurophysiology*, 111(6), 1165-1182.
- Berniker, M., O'Brien, M. K., Kording, K. P., & Ahmed, A. A. (2013). An examination of the generalizability of motor costs. *PLoS ONE*, 8(1).

- Binz, M., Dasgupta, I., Jagadish, A. K., Botvinick, M., Wang, J. X., & Schulz, E. (2023). Meta-learned models of cognition. *Behavioral and Brain Sciences*, 1-38.
- Boettcher, S. E., Gresch, D., Nobre, A. C., & van Ede, F. (2021). Output planning at the input stage in visual working memory. *Science Advances*, 7(13), eabe8212.
- Bond, K. M., & Taylor, J. A. (2017). Structural learning in a visuomotor adaptation task is explicitly accessible. *eNeuro*, 4(4).
- Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature neuroscience*, 20(7), 997-1003.
- Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, 23(5), 408-422.
- Braun, D. A., Waldert, S., Aertsen, A., Wolpert, D. M., & Mehring, C. (2010). Structure learning in a sensorimotor association task. *PLoS ONE*, 5(1).
- Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, 6(8), 1112-1125.
- Catalano, J. F., & Kleiner, B. M. (1984). Distant transfer in coincident timing as a function of variability of practice. *Perceptual and Motor Skills*, 58(3), 851-856.
- Colas, J. T., Dundon, N. M., Gerraty, R. T., Saragosa-Harris, N. M., Szymula, K. P., Tanwisuth, K., et al. (2022). Reinforcement learning with associative or discriminative generalization across states and actions: fMRI at 3 T and 7 T. *Human Brain Mapping*, 43(15), 4750-4779.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35(7), 1024-1035.

Correa, C. G., Ho, M. K., Callaway, F., Daw, N. D., & Griffiths, T. L. (2023). Humans decompose tasks by trading off utility and computational cost. *PLoS Computational Biology*, 19(6), e1011087.

Cullen, K. E., & Brooks, J. X. (2015). Neural correlates of sensory prediction errors in monkeys: Evidence for internal models of voluntary self-motion in the cerebellum. *Cerebellum*, 14, 31-34.

Curran, T., & Keele, S. W. (1993). Attentional and nonattentional forms of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(1), 189.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning: Attention and performance XXIII* (pp. 3-38). Oxford University Press.

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.

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.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 308(1135), 67-78.

Du, Y., Krakauer, J. W., & Haith, A. M. (2022). The relationship between habits and motor skills in humans. *Trends in Cognitive Sciences*, 26(5), 371-387.

- Dundon, N. M., Colas, J. T., Garrett, N., Babenko, V., Rizor, E., & Yang, D. et al. (2022). Decision heuristics in contexts exploiting intrinsic skill. *bioRxiv*.
- Elsner, B., & Hommel, B. (2004). Contiguity and contingency in action-effect learning. *Psychological Research*, 68, 138-154.
- Éltető, N., & Dayan, P. (2023). Habits of mind: Reusing action sequences for efficient planning. *arXiv preprint arXiv:2306.05298*.
- Estes, W. K., & Burke, C. J. (1953). A theory of stimulus variability in learning. *Psychological Review*, 60(4), 276.
- Fermin, A. S., 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. S., 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. M., & Posner, M. I. (1967). *Human performance*. Brooks/Cole.
- Franklin, N. T., & Frank, M. J. (2018). Compositional clustering in task structure learning. *PLoS Computational Biology*, 14(4), e1006116.
- 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.

Gershman, S. J. (2017). Context-dependent learning and causal structure. *Psychonomic bulletin & review*, 24, 557-565.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, 117(1), 197.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4), 585-595.

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*.

Hadjiosif, A. M., Krakauer, J. W., & Haith, A. M. (2021). Did we get sensorimotor adaptation wrong? Implicit adaptation as direct policy updating rather than forward-model-based learning. *Journal of Neuroscience*, 41(12), 2747-2761.

Haith, A. M., & Krakauer, J. W. (2013). Model-based and model-free mechanisms of human motor learning. In *Progress in Motor Control: Neural, Computational and Dynamic Approaches* (pp. 1-21). Springer.

Hardwick, R. M., Forrence, A. D., Krakauer, J. W., & Haith, A. M. (2019). Time-dependent competition between goal-directed and habitual response preparation. *Nature Human Behaviour*, 3(12), 1252-1262.

Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2), 100-107.

Heald, J. B., Lengyel, M., & Wolpert, D. M. (2021). Contextual inference underlies the learning of sensorimotor repertoires. *Nature*, 600(7889), 489-493.

Hikosaka, O., Rand, M. K., Miyachi, S., & Miyashita, K. (1995). Learning of sequential movements in the monkey: process of learning and retention of memory. *Journal of Neurophysiology*, 74(4), 1652-1661.

Howard, I. S., Wolpert, D. M., & Franklin, D. W. (2013). The effect of contextual cues on the encoding of motor memories. *Journal of neurophysiology*, 109(10), 2632-2644.

Huang, S., Kanervisto, A., Raffin, A., Wang, W., Ontañón, S., & Dossa, R. F. J. (2022). A2C is a special case of PPO. *arXiv preprint arXiv:2205.09123*.

Hunt, L. T., Daw, N. D., Kaanders, P., MacIver, M. A., Mugan, U., Procyk, E., Redish, A. D., Russo, E., Scholl, J., Stachenfeld, K., & Wilson, C. R. (2021). Formalizing planning and information search in naturalistic decision-making. *Nature Neuroscience*, 24(8), 1051-1064.

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.

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

Jax, S. A., & Rosenbaum, D. A. (2007). Hand path priming in manual obstacle avoidance: evidence that the dorsal stream does not only control visually guided actions in real time. *Journal of Experimental Psychology: Human Perception and Performance*, 33(2), 425.

Jensen, K. T., Hennequin, G., & Mattar, M. G. (2024). A recurrent network model of planning explains hippocampal replay and human behavior. *Nature Neuroscience*.

Jiménez, L., Vaquero, J. M., & Lupiáñez, J. (2006). Qualitative differences between implicit and explicit sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3), 475.

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.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16(3), 307-354.

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.

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.

Kerr, R., & Booth, B. (1978). Specific and varied practice of motor skill. *Perceptual and Motor Skills*, 46(2), 395-401.

Kim, D., Park, G. Y., O'Doherty, J. P., & Lee, S. W. (2019). Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. *Nature Communications*, 10(1), 5738.

- Klein, J., Spencer, S. J., & Reinkensmeyer, D. J. (2012). Breaking it down is better: haptic decomposition of complex movements aids in robot-assisted motor learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(3), 268-275.
- Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, 28(9), 1321-1333.
- 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.
- Krakauer, J. W., Hadjiosif, A. M., Xu, J., Wong, A. L., & Haith, A. M. (2019). Motor learning. *Comprehensive Physiology*, 9(2), 613-663.
- Krakauer, J. W., Pine, Z. M., Ghilardi, M. F., & Ghez, C. (2000). Learning of visuomotor transformations for vectorial planning of reaching trajectories. *Journal of Neuroscience*, 20(23), 8916-8924.
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 249.
- Krusche, M. J., Schulz, E., Guez, A., & Speekenbrink, M. (2018). Adaptive planning in human search. *bioRxiv*, 268938.
- Kumar, S., Dasgupta, I., Cohen, J. D., Daw, N. D., & Griffiths, T. L. (2020). Meta-learning of structured task distributions in humans and machines. *arXiv preprint arXiv:2010.02317*
- Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3), 687-699.

- Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43, e1
- Liu, X., Mosier, K. M., Mussa-Ivaldi, F. A., Casadio, M., & Scheidt, R. A. (2011). Reorganization of finger coordination patterns during adaptation to rotation and scaling of a newly learned sensorimotor transformation. *Journal of Neurophysiology*, 105(1), 454-473.
- Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95(4), 492.
- Martin, T. A., Keating, J. G., Goodkin, H. P., Bastian, A. J., & Thach, W. T. (1996). Throwing while looking through prisms: I. Focal olivocerebellar lesions impair adaptation. *Brain*, 119(4), 1183-1198.
- Mattar, A. A., & Ostry, D. J. (2010). Generalization of dynamics learning across changes in movement amplitude. *Journal of Neurophysiology*, 104(1), 426-438.
- Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 21(11), 1609-1617.
- Mattar, M. G., & Lengyel, M. (2022). Planning in the brain. *Neuron*, 110(6), 914-934.
- McCracken, H. D., & Stelmach, G. E. (1977). A test of the schema theory of discrete motor learning. *Journal of Motor Behavior*, 9(3), 193-201.
- McDougle, S. D., Bond, K. M., & Taylor, J. A. (2015). Explicit and implicit processes constitute the fast and slow processes of sensorimotor learning. *Journal of Neuroscience*, 35(26), 9568-9579.

- McDougle, S. D., Wilterson, S. A., Turk-Browne, N. B., & Taylor, J. A. (2022). Revisiting the role of the medial temporal lobe in motor learning. *Journal of Cognitive Neuroscience*, 34(3), 532-549.
- McNamee, D., & Wolpert, D. M. (2019). Internal models in biological control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1), 339-364.
- Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks*, 9(8), 1265-1279.
- Miller, K. J., Botvinick, M. M., & Brody, C. D. (2018). From predictive models to cognitive models: An analysis of rat behavior in the two-armed bandit task. *bioRxiv*.
- Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review*, 126(2), 292.
- Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, 117(3), 363.
- Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning* (pp. 1928-1937). PMLR.
- Moore, E. F. (1959). The shortest path through a maze. In *Proc. of the International Symposium on the Theory of Switching* (pp. 285-292). Harvard University Press.
- Mosier, K. M., Scheidt, R. A., Acosta, S., & Mussa-Ivaldi, F. A. (2005). Remapping hand movements in a novel geometrical environment. *Journal of Neurophysiology*, 94(6), 4362-4372.

- Mussa-Ivaldi, F. A. (1999). Modular features of motor control and learning. *Current Opinion in Neurobiology*, 9(6), 713-717.
- Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, 25(9), 757-775
- Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, 30(37), 12366-12378.
- Newell, K. M. (1985). Coordination, control and skill. In D. Goodman & I. M. Franks (Eds.), *Advances in psychology* (pp. 295-317). North-Holland.
- Newell, K. M. (1991). Motor skill acquisition. *Annual Review of Psychology*, 42, 213-237.
- Newell, K. M., & Shapiro, D. C. (1976). Variability of practice and transfer of training: Some evidence toward a schema view of motor learning. *Journal of Motor Behavior*, 8(3), 233-243.
- Nilsson, N. J. (1980). *Principles of Artificial Intelligence*.
- Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, 19(1), 1-32.
- Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, 12(7), 265-272.
- O'keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford university press.
- Park, J. H., Wilde, H., & Shea, C. H. (2004). Part-whole practice of movement sequences. *Journal of Motor Behavior*, 36(1), 51-61.

- Petrides, M. (1997). Visuo-motor conditional associative learning after frontal and temporal lesions in the human brain. *Neuropsychologia*, 35(7), 989-997.
- Pu, Y., Kong, X. Z., Ranganath, C., & Melloni, L. (2022). Event boundaries shape temporal organization of memory by resetting temporal context. *Nature communications*, 13(1), 622.
- Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268), 1-8.
- Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*.
- Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, 17(3), 147-159.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99). New York: Appleton-Century-Crofts.
- Rios, L. H., & Chaimowicz, L. (2010). A survey and classification of A\* based best-first heuristic search algorithms. In *Brazilian Symposium on Artificial Intelligence* (pp. 253-262). Berlin, Heidelberg: Springer.
- RStudio Team. (2023). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA. Retrieved from <http://www.rstudio.com/>
- Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, 82(4), 225.

- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593-1599.
- Schvaneveldt, R. W., & Gomez, R. L. (1998). Attention and probabilistic sequence learning. *Psychological Research*, 61(3), 175-190.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461-464.
- Shadmehr, R. (2004). Generalization as a behavioral window to the neural mechanisms of learning internal models. *Human Movement Science*, 23(5), 543-568.
- Shadmehr, R., & Mussa-Ivaldi, F. A. (1994). Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience*, 14(5), 3208-3224.
- Shadmehr, R., Smith, M. A., & Krakauer, J. W. (2010). Error correction, sensory prediction, and adaptation in motor control. *Annual Review of Neuroscience*, 33, 89-108.
- Shen, S., & Ma, W. J. (2016). A detailed comparison of optimality and simplicity in perceptual decision making. *Psychological Review*, 123(4), 452.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., & Chen, Y. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354-359.
- Simon, D. A., & Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision task in humans. *Journal of Neuroscience*, 31(14), 5526-5539.

Snider, J., Lee, D., Poizner, H., & Gepshtein, S. (2015). Prospective optimization with limited resources. *PLoS Computational Biology*, 11(9), e1004501.

Soch, J., & Allefeld, C. (2016). Exceedance probabilities for the Dirich

Sokolov, A. A., Miall, R. C., & Ivry, R. B. (2017). The cerebellum: Adaptive prediction for movement and cognition. *Trends in Cognitive Sciences*, 21(5), 313-332.

Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, 112(37), 11708-11713.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. MIT Press.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, 46(4), 1004-1017.

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of learning and memory*, 82(3), 171-177.

Taylor, J. A., Klemfuss, N. M., & Ivry, R. B. (2010). An explicit strategy prevails when the cerebellum fails to compute movement errors. *The Cerebellum*, 9, 580-586.

Taylor, J. A., Krakauer, J. W., & Ivry, R. B. (2014). Explicit and implicit contributions to learning in a sensorimotor adaptation task. *Journal of Neuroscience*, 34(8), 3023-3032.

The MathWorks Inc. (2022). *MATLAB version: 9.13.0 (R2022b)*. Natick, Massachusetts: The MathWorks Inc. Retrieved from <https://www.mathworks.com>

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Macmillan Press.

Thorndike, E. L. (1927). The law of effect. *American Journal of Psychology*, 39(1/4), 212-222.

Thoroughman, K. A., & Taylor, J. A. (2005). Rapid reshaping of human motor generalization. *Journal of Neuroscience*, 25(39), 8948-8953.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189.

Unterrainer, J. M., & Owen, A. M. (2006). Planning and problem solving: from neuropsychology to functional neuroimaging. *Journal of Physiology-Paris*, 99(4-6), 308-317.

van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2017). A computational model for decision tree search. *CogSci*.

van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, 618(7967), 1000-1005.

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

Velázquez-Vargas, C. A., Daw, N. D., & Taylor, J. A. (2023). Learning generalizable visuomotor mappings for de novo skills. *bioRxiv*, 2023-07.

Velazquez-Vargas, C. A., & Taylor, J. (2023). Exploring human learning and planning in grid navigation with arbitrary mappings. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45, p. 45).

Velázquez-Vargas, C. A., & Taylor, J. A. (2024). Working memory constraints for visuomotor retrieval strategies. *Journal of Neurophysiology*.

Verwey, W. B. (2001). Concatenating familiar movement sequences: The versatile cognitive processor. *Acta Psychologica*, 106(1-2), 69-95.

- Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., ... & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6), 860-868.
- Welch, B. L. (1947). The generalization of 'STUDENT'S' problem when several different population variances are involved. *Biometrika*, 34(1-2), 28-35.
- Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(6), 1047.
- Wilson, R. C., Nassar, M. R., & Gold, J. I. (2013). A mixture of delta-rules approximation to Bayesian inference in change-point problems. *PLoS Computational Biology*, 9(7).
- Wilterson, S. A., & Taylor, J. A. (2021). Implicit visuomotor adaptation remains limited after several days of training. *eneuro*, 8(4).
- Wise, S. P., & Murray, E. A. (2000). Arbitrary associations between antecedents and actions. *Trends in Neurosciences*, 23(6), 271-276.
- Yang, C. S., Cowan, N. J., & Haith, A. M. (2021). De novo learning versus adaptation of continuous control in a manual tracking task. *eLife*, 10.
- Ying, Z., Callaway, F., Kiyonaga, A., & Mattar, M. G. (2023). Resource-rational encoding of reward information in planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).
- Yoo, A. H., & Collins, A. G. (2022). How working memory and reinforcement learning are intertwined: A cognitive, neural, and computational perspective. *Journal of Cognitive Neuroscience*, 34(4), 551-568.

## Publications

### Portions of this dissertation have previously been published in:

Velázquez-Vargas, C. A., Daw, N. D., & Taylor, J. A. (2024). The Role of Training Variability for Model-based and Model-free learning of an Arbitrary Visuomotor Mapping. *PLoS Computational Biology*.

Velázquez-Vargas, C. A., & Taylor, J. A. (2024). Working memory constraints for visuomotor retrieval strategies. *Journal of Neurophysiology*.

Velázquez-Vargas, C. A., Christian, I. R., Taylor, J. A., & Kumar, S. (2024). Learning to Abstract Visuomotor Mappings using Meta-Reinforcement Learning. In *Proceedings of the annual meeting of the cognitive science society* (No. 46).

Velázquez-Vargas, C. A., & Taylor, J. (2023). Exploring human learning and planning in grid navigation with arbitrary mappings. In *Proceedings of the annual meeting of the cognitive science society* (No. 45).

### Preprints:

Velázquez-Vargas, C. A., & Taylor, J. A. (2024). Learning to Move and Plan like the Knight: Sequential Decision Making with a Novel Motor Mapping. *bioRxiv*.