Eye-movements as a signature of age-related differences in global planning strategies for spatial navigation

Elisa M. Tartaglia, Celine Boucly, Guillaume Tatur, Angelo Arleo

Eye-movements as a signature of age-related differences in global planning strategies for spatial navigation

Elisa M. Tartaglia[1,2], Celine Boucly[1], Guillaume Tatur[1], and Angelo Arleo[1]

[1]Sorbonne University, UPMC University of Paris 06, INSERM, CNRS, Vision Institute, F-75012 Paris, France.
[2]Essilor International, R&D, Vision Science, Paris, France.

November 28, 2018

Abstract

The ability to efficiently find alternatives routes when faced with unexpected obstacles along our path is among the most compelling evidence of the flexibility of human behaviour. Although a plethora of plausible computations have been put forward to elucidate how the brain accomplishes efficient goal-oriented navigation, the mechanisms that guide an effective re-planning when facing obstructions are still largely undetermined. There is a fair consensus in postulating that possible alternatives routes are internally replayed sampling from past experiences, however, there is currently no account of the criterion according to which those memories are replayed. Here, we posit that paths, which are expected to be more rewarding are replayed more often and that eye movements are the explicit manifestation of this re-planning strategy. In other words, the visual sampling statistics reflects the retrieval of available routes on a mental representation of the environment.
To test our hypothesis, we measured the ability of both young and old human subjects to solve a virtual version of the Tolman maze, while we recorded their eye movements. We used reinforcement learning (RL) to corroborate that eye movements statistics was crucially subtending the decision making process involved in re-planning and that the incorporation of this additional information to the algorithm was necessary to reproduce the behavioral performance of both screened populations.

Introduction

The ability to efficiently find alternatives routes when faced with unexpected obstacles along our path is among the most compelling evidence of the flexibility of human behaviour. If, for example, our usual way to the bakery is blocked by some construction work, we would still be able to reach our destination by making -hopefully- a small detour.
Although a plethora of plausible computations have been put forward to elucidate how

the brain accomplishes efficient goal-oriented navigation (for a review see Madl et al. 2015; Chersi and Burgess 2015), the mechanisms that guide an effective re-planning when facing obstructions are still largely undetermined.

There is a fair consensus in postulating that possible alternatives routes are internally replayed sampling from past experiences via network interactions involving hippocampus and prefrontal cortex (Spiers and Gilbert 2015). However, there is currently no account of the criterion according to which those memories are replayed. Here, we posit that paths which are expected to be more rewarding are replayed more often and that eye movements are the explicit manifestation of this re-planning strategy. In other words, the visual sampling statistics reflects the retrieval of available routes on a mental representation of the environment (i.e. on a cognitive map).

To test our hypothesis, we measured the ability of human subjects to solve a virtual version of the Tolman maze, while we recorded their eye movements. In his original study, Tolman found that rodents were capable to take detours and to select the shortest path to food pellets. Consequently, he put forward that an internal representation of the physical space is a necessary condition to be able to re-plan one's way in a modified environment (Tolman 1948). In the light of these results, we expected young adults not only to be able to successfully re-plan their way through the maze, but also to be able to sketch a faithful representation of the maze, as further evidence of a cognitive map formation (one that cannot be accounted for in non-human subjects). According to our hypothesis, then, the eye movements statistics collected when subjects are required to find alternative routes, should reflect planning over the cognitive map; if such map is correct, the correct alternative route will be chosen.

A much more challenging instance to test our hypothesis upon concerns older adults performance. Aside from general spatial navigation deficits (Barrash 1994; Wilkniss et al. 1997; Burns 1999; Newman and Kaszniak 2000; Moffat and Resnick 2002; Driscoll et al. 2005; Head and Isom 2010; Jansen et al. 2010; Lester et al. 2017), aging has been specifically linked to a delay in cognitive map formation as well as to an impediment in using it (Iaria et al. 2009). Moreover, elderly have been shown to be defective in allocentric navigation (Moffat et al. 2006; Antonova et al. 2009; Rodgers et al. 2012; Bohbot et al. 2012; Gazova et al. 2013), which would suggest that they rely more on a stimulus-response association type of navigation strategy, rather than on planning. According to our hypothesis, instead, older adults, as their younger counterpart, are capable to do mental planning; likewise, their eye movements reflect planning over a cognitive map. However, we conjectured, older adults re-plan their alternative route over a degraded cognitive map; this might explain why older subjects have been found to be impeded in solving a detour task and/or in taking shortcuts (Harris and Wolbers 2014).

To corroborate that eye movements statistics was crucially subtending the decision making process involved in re-planning, we used reinforcement learning (RL) models. Although a phenomenological model, both neuroimaging and neurophysiological studies have provided strong indication that specific anatomically identified brain regions are indeed performing RL based computations (Schultz et al. 1997; Berns et al. 2001; O'Doherty 2004; Pagnoni et al. 2002). Moreover, the use of RL to fit behavioural data of (young) observers performing a navigation task has provided a clear indication of the strategies at use. In particular, contrasting model-based and model-free RL algorithms have led to identify whether navigation strategies involve or not the ability of planning

future path selection (Daw et al. 2005; Simon and Daw 2011; Keramati et al. 2011; Gershman et al. 2014; Tartaglia et al. 2017).

Here, we show that the incorporation of EM statistics as additional information to RL algorithms was necessary to reproduce the behavioral performance of both screened populations. In the classical RL theory, the *agent* (e.g. the observer navigating through the maze) learns from reinforcement the future value of the expected reward at any given *state* of the environment (e.g. a location in the maze). The value of a state is updated only if the agent has visited it (Sutton and Barto 1998). Here, we made a much stronger assumption: the states that observers looked at are used for planning and consequently updated even if they have not been visited. The rationale behind this assumption goes as follow: often, in a natural setting, visual exploration of the environment alone provides enough information about which direction should be taken next. Imagine you are at a crossroad, looking to the right you realize the street is dead-ended, hence you plan to turn left instead. In classical RL, the state values on the right would not get updated since the agent did not visit them. However, we argued, a more efficient and biologically plausible algorithm should be able to assign (in this case negative) values to the right path.

More into details, using the recorded data, we fit our RL algorithm to the behavioural choices (i.e. the discretised sequence of maze locations visited by each observer), as well as to the number of fixations performed during the test trials by both young and old observers. In particular, as a first step, we fed to the model (with two free parameters) the sequence of states visited during the exploration phase by each individual subject. This method provided individual (and more realistic) initial conditions to be used, in a second step, in the test trials, during which we compared observers' performance to the model performance. Importantly, the model was enriched with a crucial information: the statistics of eye movements made by each individual subject when planning their next move towards the goal. As a final step, we determined the region of parameters which yielded the closest performance to the behavioral data.

We seek to provide, for the first time, a plausible mechanistic account of the process of planning alternative routes when facing unexpected obstacles. Furthermore, we expect these results to lead to a quantification of the extent to which such process differ from young to older observers.

Methods

Participants and set-up

We tested $n = 39$ participants. Age and gender of the sample are reported in the following table.

Gender	Young (< 45 years)	Old (>65 years)
Female	9	10
Male	10	10

The mean age for young and old observers was, respectively, $27.3y$ and $78.5y$.

Experiments were run on a simulation module developed in Unity 3D, which allowed to create the virtual maze and to record subjects trajectories. A view from above of the maze, together with its dimensions is shown in Fig 1A. A chest at the end of the

maze was used as Goal (Fig 1B). Any visual cue within the environment, which could help subjects to reach the Goal faster, was avoided.
Three main paths allowed to reach the chest: a central corridor, a left and a right path, rejoining the central corridor at different locations (exit points). The central corridor provided the shortest path to the Goal, while the right one provided the longest (Fig 1A).
Exit points of both right and left paths were hidden by virtual doors, to prevent participants to get a glimpse of the maze structure from the very starting point; virtual doors opened only when subjects were in their close proximity and were not identifiable otherwise.
Subjects were generally allowed to take any path, from any direction, within the maze (except when specified otherwise).
The virtual maze was presented on a projection screen ($0.9m * 1.8m$). Subjects were sitting at a distance of $1.5m$ from the screen and moved within the environment -from a first person perspective- using a joystick. Turning rate and walking speed were kept constant i.e. all participants navigated within the environment at the same speed. This allowed to minimize differences in performance that might arise from age-related difficulties in using the joystick. We synchronized the virtual walking with footsteps sound, to render the virtual experience more realistic.
Binocular eye movements were recorded throughout the experiment with a head-mounted eye tracking system (Eyelink II, SR Research). The calibration of the eye tracking was done before the experiment and was regularly checked upon during the whole experimental session.

Procedure

Before giving their consent, subjects were informed that they were going to participate, for about 1.5 hours, in a virtual navigation study, in which their eye movements would have been recorded. The experiment was divided into three main phases: *training*, *exploration* and *testing*.
During training, instructions were verbally given and assistance with joystick manipulation and task understanding was constantly provided. During both exploration and testing phase, instructions appeared on the screen to minimize any experimenter bias.

Training phase. To reduce motor skills differences, we asked subjects to practice the use of the joystick while navigating in a virtual environment -which was different from the Tolman maze. Several repetitions of this training task were run, depending on how good the experimenter judged the individual ability to handle the joystick. Once subjects were judged at ease with the set-up, they were asked to practice a *pointing task* in yet another rendering of the virtual environment. In this case navigation was guided, i.e. subjects were asked to follow a given trajectory until a compass appeared on the screen. The task consisted in using the joystick to point the compass arrow towards a previously visited location, which was out of sight from the current position. The rationale of this phase was to get the subjects acquainted to the use of the compass in the virtual environment, before the proper pointing task in the Tolman maze was carried out. Being able to point to the correct direction provided an indication of the observers ability to locate themselves in the environment, as well as an indirect

measure of their ability to form an internal representation of the environment. Eye movements were not recorded during the training phase.

Exploration phase. Subjects were first instructed to freely navigate in the virtual Tolman maze, of whom they ignored the structure; we informed them that a chest was located somewhere in the maze, although they were not explicitly instructed to look for it, or to memorize its location. Subjects were also aware that the chest acted like a portal: once reached it, they would be instantaneously brought back to the maze starting point. In other words, the chest allowed to resume navigation from a fixed reference point, providing a useful direction cue. This first stage of exploration lasted six minutes.
Next, we instructed subjects to follow popping out arrows guiding them from the starting state to the Goal trough the three main maze paths, i.e the left, the central and the right path. This "guided exploration" phase lasted not more than two minutes and was designed to ensure that all subject went through each of the three paths at least once before the test phase.
A second session of free exploration, identical to the first one, followed.
The whole exploration phase lasted twelve minutes (such duration was determined based on a previous pilot study). Eye movements were recorded all along.

Testing phase. We used three different kind of post-exploration tests to evaluate subjects performance, i.e. their ability to grasp the Tolman maze structure. Each test phase will be described in details in the following section.

Performance measurements

Detour task

Participants were instructed to reach the chest from the starting point by traveling the shortest possible distance. The structure of the maze as well as the chest location were left unchanged with respect to the exploration phase. Observers performed a sequence of 7 trials, which we labelled as: $0 - A - B - A - B - A - B$. In condition 0, all paths were available, hence the central corridor provided the shortest path to the chest; in condition A, a portion of the central corridor was blocked, therefore the left path became the shortest; in condition B, a portion of both the central corridor and the left path were blocked and the right path became the shortest way to the Goal (Fig 1A). Eye movements were recorded during the whole session.
To quantify performance we measured for each subject i, each trial repetition j ($j = 1, 2, 3$) and each trial type k ($k = 0, A, B$), the length of the distance travelled from Start to Goal, D_{ijk}, to which we subtracted the shortest path length $min(D_k)$, i.e. the length of the shortest possible path leading from Start to Goal:

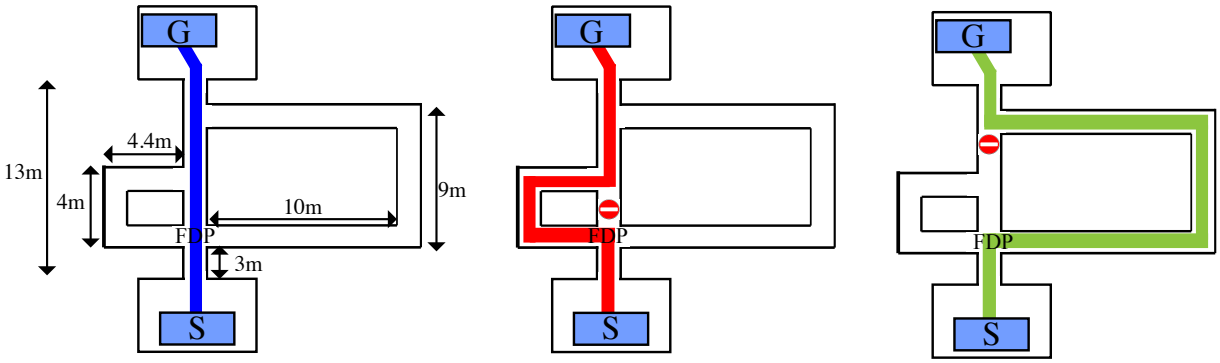
$$\Delta d_{ijk} = D_{ijk} - min(D_k) \tag{1}$$

Given the above definition, $\Delta d_{ijk} \in [0, \infty[$; performance was optimal when $\Delta d_{ijk} = 0$.

Sketched map task

After a quick break, participants were instructed to draw on an electronic tablet a bird's eye view of the maze. They were asked to label the hallways, the obstacles

A.



B.

C.

Figure 1: **A.** Shortest paths from Start (labelled as S) to Goal (labelled as G) in the three conditions of the test trials: from left to right, no block (0), block A, block B respectively. FDP labels the first decision point in the maze. The actual dimensions of the maze are also indicated. **B.** A picture of the chest located in G. **C.** Simulator view of the maze from the starting point S in the subjects' prospective.

they encountered and the chest location. They could take as long as they needed to complete the task. Eye movements were not recorded.
To quantify performance we designed a grading system and asked four unbiased raters to grade the subjects' maps according to several criteria (see Table in Appendix and Billinghurst and Weghorst 1995). A maximum of 20 points was assigned if the maze was correctly drawn. Negative points were assigned when subjects drawn extra-elements, i.e. hallways or intersections that did not exist (see Fig. 4C).

Pointing task

After sketching the map, participants were presented with the virtual Tolman maze again and were asked to point toward the chest in four different locations along the left and the right path (Fig. 3). We measured performance by calculating the angular deviation of each subject choice from the actual Goal location, i.e. the error subjects made in pointing towards the chest from their current position.

Eye-movements analysis

Eye movements were recorded binocularly at a rate of $250Hz$ during both the exploratory and the testing phase. The EyeLink II system allowed to detect saccades and blinks by setting threshold parameters for saccade's velocity ($30°/s$), acceleration ($9000°/s^2$) and motion ($0.15°$). Periods between saccades in which blinks were absent, were identified as fixations. A fixation was quantified in terms of *gaze direction*, i.e. the angle between the centre of the virtual camera (which coincides with the position of the subject in the virtual space) and the 2D fixation point on the screen, whose coordinates were directly provided by the eye tracking device. From the point of view of a subject moving in the virtual space, the horizontal field of view spanned approximately 106°.
For the analysis, we focused on the subset of gaze directions recorded while subjects navigated from the maze starting point to the first decision point (*FDP*, see Fig. 1A), i.e. the portion of space in which subjects were expected to make a decision about which of the three paths to choose. This portion of the maze, which we referred to as *region of interest*, spanned $5.5m$. We analyzed, separately, gaze directions recorded during the exploration phase and those recorded during each test trial, pooling together gaze directions for each population and test trials of the same kind.
We were interested in testing two hypothesis concerning eye movements dynamics: first, as a manifestation of learning, the pattern of gaze directions changed significantly when comparing the exploration phase to the test trials; second, as a manifestation of a different decision making dynamics, the pattern of gaze directions within test trials changed significantly when comparing younger to older adults. To test our hypothesis, we partitioned the region of interest in several subregions, for each of which we built a gaze direction distribution polling together the gaze direction data of the whole population. The number of subregions was determined by a sliding window of size $0.5m$ and step $0.05m$. Such a representation allowed us to assess and to compare the evolution in time, from start to *FDP*, of the gaze direction distribution of each population.
To be able to incorporate eye movements during re-planning of an alternative route in the model, we analysed eye movements statistics for each individual observer during test trials. In particular, we counted the proportion of fixations made straight ahead

towards the central corridor, towards the right and towards the left arm while the observer moved from start to *FDP*. Note that we discarded the number of times in which observer looked behind towards the starting point simply because none of the observers went back to it before reaching the *FDP*. Subsequently, we fed the proportion of eye movements to the model.

Model

We used hierarchical reinforcement learning (HRL) to model subjects'performance in the Tolman maze task. HRL, unlike standard RL algorithms, allows the learning agent to take actions extended in space and to keep track of the distance travelled while executing them. We referred to this class of actions as *options*. Formally, each option is labelled by its initial and terminal state, as well as by its length z (Fig.6A).
In a given state s, the agent chooses an option o according to the following probability rule (*soft-max policy*):

$$P(o|s) = \frac{e^{\frac{Q(s,o)}{\tau}}}{e^{\sum_{o' \in O} \frac{Q(s,o')}{\tau}}} \tag{2}$$

where $Q(s, o)$ is the value associated to option o while in s and τ is a "noise" parameter. The normalisation takes into account all available options o' in state s. If $\tau \to 0$, the option with the highest probability to be selected is the one associated with the largest Q value. In this case, the agent is said to be *greedy*. To the contrary, the larger the τ the more the agent is *explorative* and is willing to choose options whose Q value is not necessarily the largest. In the limit of $\tau \to \infty$, all available options in s become equiprobable.
Once the option o has been chosen, the Q value associated to the option initial state s is updated according to:

$$Q(s, o) = Q(s, o) + \alpha\delta \tag{3}$$

where α controls the learning rate and δ adjust the previous estimation of the reward given the reward r collected when executing o from s:

$$\delta = r + \gamma^z max_{o'} Q(s', o') - Q(s, o). \tag{4}$$

As in standard Q-learning algorithms, γ is the discount factor, which weights the rewards according to how far from s they will be obtained ($0 \leq \gamma \leq 1$). Unlike standard Q-learning, here γ is affected by the option length z, i.e. by the number of steps elapsed since o was chosen, so that longer options are discounted more.
The execution of option o brings the agent to the option terminal state s'. The task of the agent is to maximise the total reward across the environment through the update of Q values estimation. If the algorithm has converged, choosing the option associated with the largest Q value at each s, ensures the highest total reward.

In the model-free version of HRL (MF-HRL), $Q(s, o)$ gets updated online, i.e. only after the agent visits s and chooses o. In the model-based version of HRL (DQ-HRL), Q values are updated also offline, through *planning*. Planning works in the following way: upon visiting s, the transition $s \to o \to s'$ and its associated reward r are stored in a matrix M to build up a "model" of the environment. Before choosing the next option, the agent uses M to randomly sample n_{pl} states and options visited in the past

and update their Q values offline. This planning procedure provides faster convergence to the optimal Q values.

In the following, we used DQ-HRL to reproduce the behavioural data. The model has four free parameters, $\tau, \alpha, \gamma, n_{pl}$; we fixed $\gamma = 0.9$ for both young and old subjects; we assumed the remaining three parameters to differ between the two populations, but not within individual subjects.

Results

Behavioral Results

As expected, we found that young observers significantly outperformed old ones in the detour task. The histogram of Δd_{ijk} shows that the whole young population performed optimally, or nearly optimally, in all trials, i.e. choosing almost always the shortest possible path to the Goal (Fig. 2A.). A much higher variability and sub-optimality was observed for older subjects, with some trajectories deviating up to $250m$ from the shortest path length (Fig. 2B.). The difference in performance between the two populations appears even more pronounced when plotting $\langle \Delta d \rangle_i$ averaged across subjects, with young subjects performing significantly better than old ones in all but one trial (Fig. 2C.,F.). Averaging further across repetitions of the same trial type (Fig. 2D.,G.), or across all trials confounded (Fig. 2E.,H.) confirmed these results. Furthermore, we found that young observers were significantly more accurate in pointing towards the Goal location in all tested sites (Fig. 3). Old subjects appear to be less and less able to locate the Goal at each change of direction in the maze (Fig. 3B). Analogous results were found when testing pointing performance across four locations spread on the left arm, or when pointing towards the starting point rather than the Goal (results not shown).

Old and young participants map grades were also significantly different. Three examples are shown in Fig. 4. When averaging maps scores across graders, we found a mean score of $17/20 \pm 0.1$ for young, versus $5/20 \pm 1$ for old subjects (Wilcoxon rank sum test, $p < 10^{-4}$), indicating a much higher accuracy in the maze rendering of the young population. The standard deviation of the scores indicates that the evaluation of old subjects maps was, as expected, much noisier.

Finally, we checked and found that the three performance measurements were significantly correlated. (Spearman test; correlation Δd_{ijk}-pointing: $p < 10^{-6}$; correlation Δd_{ijk}-map: $p = 0.005$; correlation pointing-map: $p = 0.002$; note that the correlation is calculated using Δd_{i1B}, i.e. only the first repetition of trial B).

Taken together, theses results consistently point to a significant and detrimental effect of age on the ability to build up a coherent representation of the environment.

Eye-movements results

Having conducted the experiment in a virtual simulator provides the advantage of isolating the visual contribution to the navigation task; in the absence of proprioceptive and vestibular signals, observers could rely exclusively on their sight to collect information about the environment. Given that, we expected to find a manifestation of learning and/or of the ongoing decision making process in the dynamics of the ob-

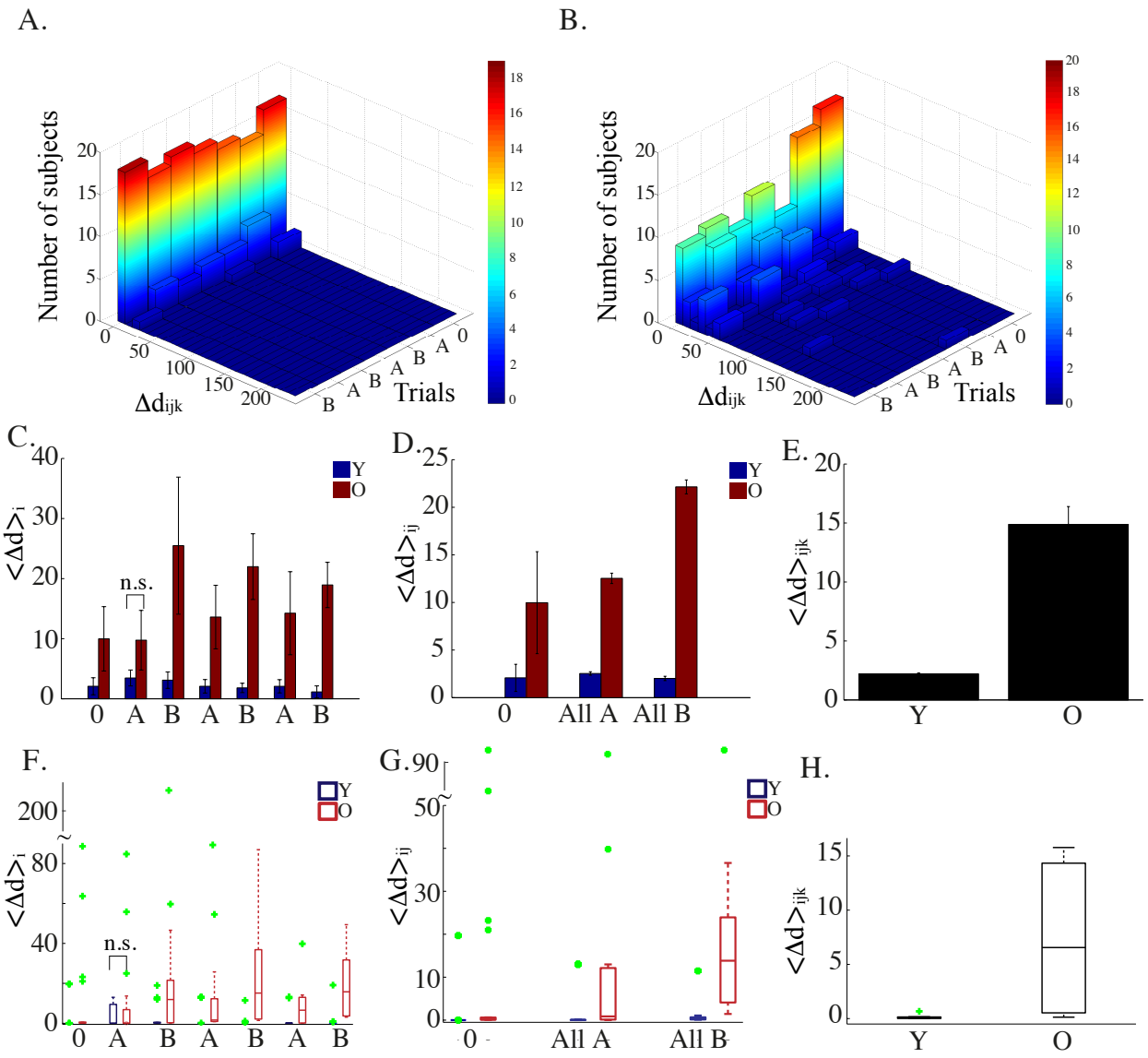


Figure 2: **Behavioral results - Detour task**. Histogram of the distance travelled from Start to Goal across the 7 trials (0-A-B-A-B-A-B) for the young (**A.**) and the old population (**B.**) An optimal performance implies $\Delta d_{i,j,k} = 0$ (see text for more details). Travelled distance averaged across observers (**C.**), across observers and trials of the same type (**D.**), across observers and all trial types (**E.**) The only non significant result is labelled as *n.s.*. All p-values have been calculated via the Wilcoxon rank sum test. No subject was excluded from the analysis. Panels **F.**, **G.**, and **H.** show the box-plots corresponding respectively to panels C., D., E. The green points label the outliers.

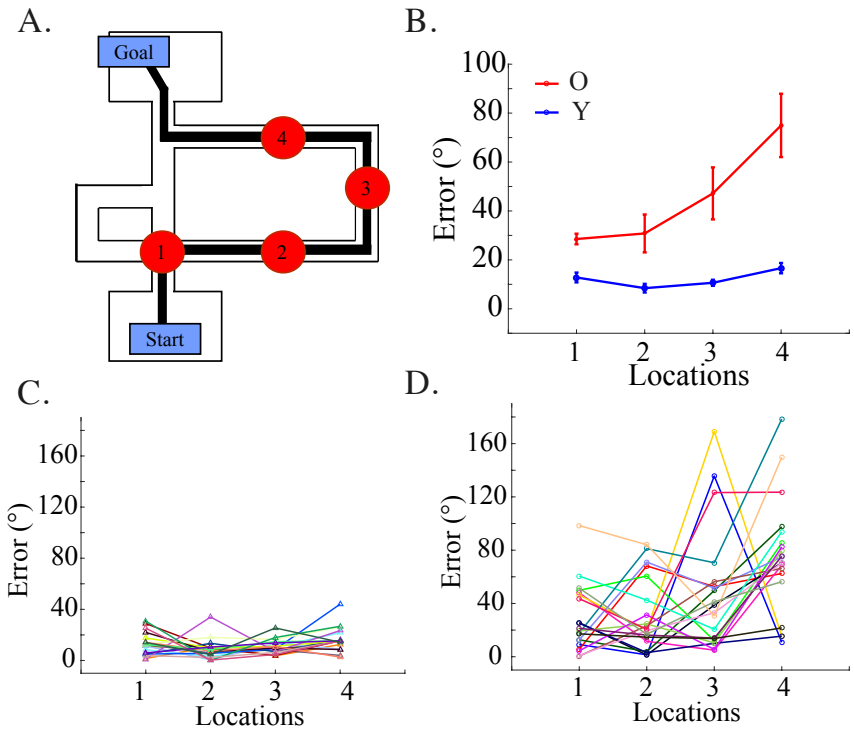


Figure 3: **Behavioral results - Pointing task. A.** The red circles label the four locations along the right path in which observers were required to stop and point to the Goal position. **B.** Average across subjects of the deviation from the Goal position measured in each of the four locations in A. While the error in the young population tends to remain constant (and very small), it increase with the number of changes in direction for the old population. Results are significant in any of the tested location (location 1, $p = 0.04$; location 2, $p = 0.001$; location 3, $p = 0.0005$; location 4, $p < 10^{-5}$; Wilcoxon rank sum test). Following the same procedure, we also tested four locations along the left path and found similar results. **C.** Same as in B., but for individual young observers. **D.** Same as in C., but for individual old observers.

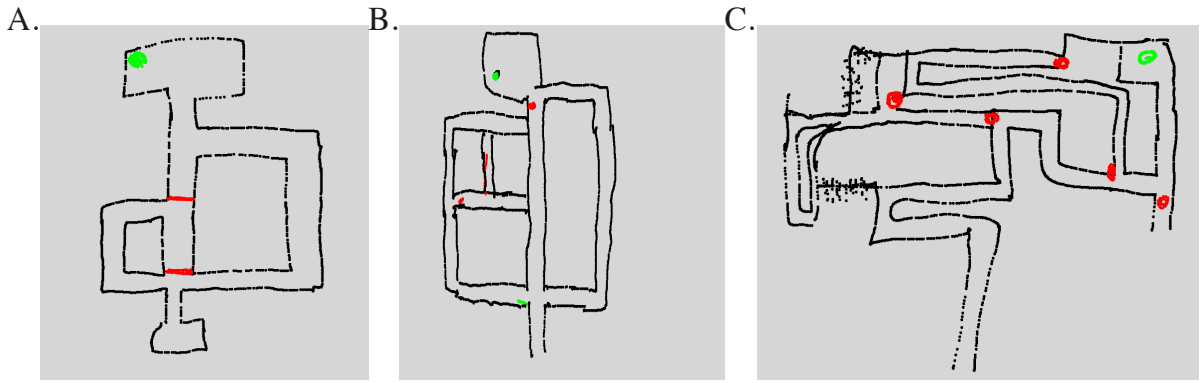


Figure 4: **Behavioral results - Sketched map task.** One young (**A.**) and two old subjects (**B., C.**) rendering of the Tolman maze. Drawing were done right after the detour task. Observers were instructed to label in green the Goal location and in red the blocks. Average grades were, from left to right, 20, 14.5 and -6.

servers' eye movements.

In particular, we anticipated significant changes in gaze direction distributions when comparing the exploration phase to the test trials within the same population. During exploration we expected gaze to be more dispersed, with participants trying to maximize the range of their eye movements to probe the unknown environment. During test trials, instead, the uncertainty should be reduced since participants had gained some insight about the environment, had a precise task to accomplish and were not expected to visually explore the environment anymore. Consequently, gaze directions distribution during test trials should be more peaked.

In accordance with our hypothesis, by comparing the gaze distibution's mean, standard deviation and kurtosis in each sliding window of the region of interest (see Procedure), we found that the gaze direction distributions of the young population was flatter during exploration and more picked during test trials. Interestingly, the opposite was observed for the elderly population (a three way ANOVA was performed with three factors: age (old/young), correctness (correct/incorrect), epoch (exploration/tests). Significant differences between exploration and trials were found for the standard deviation and the kurtosis of horizontal gaze distributions for young vs old participants; results not shown). A possible interpretation of these results could be that young participants were more exploratory at first (flat gaze distribution), while they rather exploited once learning had taken place (picked gaze distribution). To the contrary, old people appeared to be less exploratory in the beginning (picked gaze distribution). This behaviour could have led to some sort of confusion during test trials and, consequently, to a flatter gaze distribution. It is worth mentioning that, while these effects were significant for age, they were not when considering correctness as a factor.

Subsequently, we analyzed gaze direction distributions of young and old populations virtually walking trough the maze region of interest while performing the test trials. In agreement with the spontaneous anticipation of locomotor trajectory by gaze direction (Bernardin et al. 2012), we found that also virtual trajectories, or, in other words, joy-

stick turns, were anticipated by gaze direction in both populations. More importantly, we hypothesised that such a gaze anticipatory effect could be delayed in older adults, as a manifestation of a lingering onset of the decision making process. Three principal gaze directional patterns were expected close to the decision point: left, forward and right oriented gaze directions, corresponding to the three possible paths to the goal. However, if we choose to analyze, for example, B trials alone, we should expect a bias of gaze direction towards the right, at least for the population who succeeded in performing the task (i.e. who chose more often the right path). While still considering young and old data categories separately, we analyzed two datasets of gaze directions, pooling together B trials repetitions: 'All' and 'Correct' data. 'All' data refers to the use of gaze directions drawn from all subjects in the considered category, disregarding whether they succeed or not the trial, whereas 'Correct' data refers to the selection of successfully completed trials.

In order to evaluate the onset of the gaze anticipatory effect in the two populations, we evaluated the location within the region of interest ($DM^*_{Y,O}$) at which the gaze direction distribution significantly departed from unimodality (Hartigan test). As a matter of fact, as soon as subjects got closer to the FDP, their gaze direction was no longer peaked at 0°, i.e. towards the central corridor, but started to shift towards the right or the left, i.e. towards the two only available paths in B trials, depending on which direction of motion was subsequently chosen. When considering All data populations, we found, as expected, that $DM^*_Y < DM^*_O$, suggesting that young observers were faster in the decision making process. Note that we constrained the speed of the virtual navigation, which was constant and the same for all subjects, hence the result did not depend on a faster motor response in young adults. As evidence that the delayed onset of gaze anticipation was not due to the uncertainty associated to unsuccessful trials but was rather a true age effect, we found that $DM^*_Y < DM^*_O$ even when we exclusively considered successful trials (see Fig.12 in Appendix).

Our main hypothesis on eye movements pertained to their statistics, rather than to their dynamics. In particular, we expected the frequency of gaze directions towards a given path to be a proxy of the amount of reward attributed to it; in other words, we hypothesised that while planning an alternative route, observers look more often in the direction from which they expect to attain the highest reward, which, in our case, coincides with the direction that they believe leads to the Goal faster. To test our hypothesis, we included in a model-based hiererchical RL algorithm (DQ-HRL) the updating of the value of the states observers looked at while moving towards the FDP; we assumed that those are the states observers are relying on to mentally plan their detour to the Goal. In the following we describe in detail how the model works.

Model Results

We first tested the ability of HRL to learn the shortest -and hence most rewarding- path to the Goal in the Tolman maze. As expected, we found that DQ-HRL learned faster and more accurately than both MF-HRL and a standard MB algorithm, which used single-step actions instead of options (i.e. Dyna-Q; Fig. 7B). In DQ-HRL, Q values associated to the first decision point attained their optimal values already after the completion of 15 episodes (Fig. 7C., right panel). Q values in Dyna-Q also reached their asymptote quite fast, i.e. starting from 20 episodes, however, the algorithm was less accurate than HRL, since it could not disentangle most actions values (i.e.

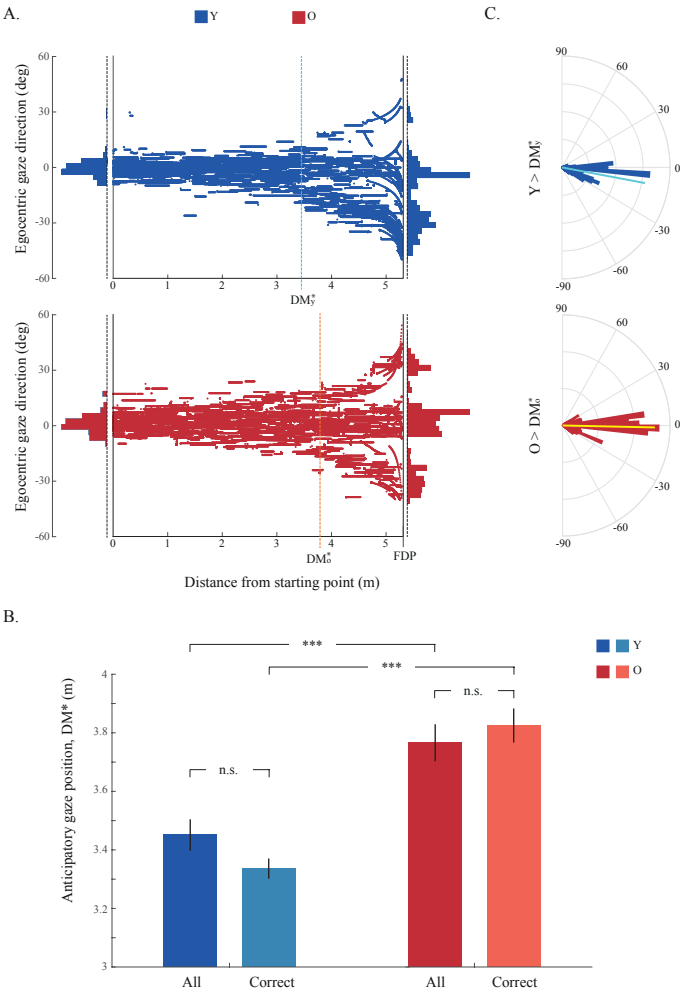


Figure ??

Figure 5: **A.** On the left plot, gaze direction distributions of 'All' Y and 'All' O data binned over a sliding window as a function of the distance between S and FDP. $DM^*_{Y,O}$ labels the location at which the distribution switches from unimodal to multimodal, i.e. the location at which the eyes anticipated the direction of motion. On the right plot, the gaze direction distribution obtained in the last bin, i.e. closest to the FDP, is represented on a semi-polar plot; the superimposed line labels the median of the distribution. Note how the gaze distribution of Y observers is much more biased towards the right direction. **B.** To estimate the variability associated to the decision making point $DM^*_{Y,O}$, we applied for each of the four datasets Young correct (YC)/old correct (OC), Young All data (YAD) / old All data (OAD) the following procedure. For each sliding window, we generated a kernel probability density estimate for the current gaze direction distribution, using a normal kernel smoothing function. We then randomly sampled 20 synthetic datasets from the pdf of the kernel distribution. For each dataset, we performed the Hartigan test. Parametric t-tests showed that 'All' Y and 'All' O data exhibited a significant ($p < 0.001$) difference between onsets of the multimodal gaze direction pattern. Within each of these groups, changes in onset value when considering 'All' or only 'Correct' trials were not significant.

$Q(L) = Q(D) = Q(R)$ in Fig. 7C., left panel). The model free version of HRL was the slowest and less accurate (Fig. 7C., central panel).
Importantly, the way in which we defined options in HRL algorithms impedes the agent to repeatedly go back and forth in small subregions of the maze, i.e. to get jammed in unrealistic loops, which we never observed in the subjects behaviour. Those unrealistic loops are hardly avoidable in standard MB algorithms, which use primitive actions.

Subsequently, we used DQ-HRL in a two-steps procedure to reproduce the performance of young (Y) and old (O) subjects during the execution of the test trials (Fig. 6B). First, we fed the model with the actual sequence of states and options taken by each subject during the *exploration* phase. Initial Q values were set to zero for all subjects; the values of α and n_{pl} were fixed and assumed to be equal for all subjects within a population; the noise parameter τ was unnecessary since state transitions in the model were not governed by any stochastic policy, but were deterministically defined by the actual trajectory followed by the subjects. The reward r associated with the very last transition bringing the agent to the Goal was set to one; all the other rewards were set to zero.
The feeding procedure provided, for each subject, the Q values at the end of the exploration phase, i.e. $Q_{0,i}$ (Fig. 6B).
The following step of the procedure consisted in *i)* setting individual $Q_{0,i}$ as initial conditions for running the model with individual *test trials*, and *ii)* using subjects gaze directions during the execution of test trials as a tool to re-plan a path within the maze. The parameters α and n_{pl} were assumed to remain unchanged for both populations with respect to the exploration phase; furthermore, we set $\tau \simeq 0$ since, contrary to the exploration phase, during the tests trials we asked subjects to be "greedy" , i.e. to reach the Goal via the shortest possible path.
We run the model for each individual subject in the seven test conditions $0 - A - B - A - B - A - B$.

To clarify the role of eye movements (EM) in the model, let us describe the procedure we followed for a given subject i engaged, for example, in a B trial. Q values were initialised at $Q_{0,i}$. Starting from $s = S$, the agent chose an option according to the policy in Eq.2, where $Q(s, o) = Q_{0,i}(s, o)$, and landed in $s' = FDP$, i.e. in the maze first decision point. The value of $Q_{0,i}(s, o)$ was updated according to Eq. 3 and the state transition was stored in M together with its associated reward r:

$$M(s, o) \leftarrow (FDP, r = 0); \tag{5}$$

we assumed that, before choosing any other option, the agent used subject i eye movements for planning his next move. From the eye-tracking data, we extracted the number and the direction of fixations made by subject i while navigating from S to FDP (Fig. 5). If, for example, 70% of the fixations were made straight ahead towards the corridor, 20% towards the right arm and, 10% towards the left arm, we stored in M the following transitions:

$$M(FDP, o_U) \leftarrow (s_U, r = r_U) \tag{6}$$
$$M(FDP, o_R) \leftarrow (s_R, r = r_R) \tag{7}$$
$$M(FDP, o_L) \leftarrow (s_L, r = r_L) \tag{8}$$

where o_U labelled the option straight ahead (or UP) and analogously, o_R (o_L) the option right (left). The state s_U labelled the terminal state of o_U coming from FDP

(analogously for s_R and s_L; Fig. 6C). Importantly, the value of the reward associated to each transition was set equal to the proportion of EM made in the corresponding direction; in our example, $r_U = -0.7$, $r_R = 0.2$ and $r_L = 0.1$. Note that we set r_U negative for all trials in which a block was used (i.e. A and B trials); we reasoned that, since the block was made clearly visible, subjects were well aware of the impossibility to choose o_U in those cases (at the expenses of bumping straight into the block) and internally associated it with a negative reward.
While still in FDP, the agent randomly sampled n_{pl} states and options from M and updated the corresponding Q values. Crucially, in our procedure EM allow the agent to do planning using a larger sample of transitions (i.e. those associated to visited states, as well as those associated to states which were only looked at) and, eventually, end up with a better and faster estimation of Q values; without EM, instead, the agent would rely only on a single transition (Eq. 5) for planning while in FDP.
When the n_{pl} number of planning steps were over, the agent chose almost always (since $\tau \simeq 0$) the option associated with the highest Q value, o_R in our example, ending up in s_R. Once the agent went past FDP, we reasoned that EM would no longer reflect a planning strategy to reach the Goal and hence we did not consider subjects EM (i.e. beyond FDP, only those transitions which were actually executed were stored in M and used for planning, as in standard MB frameworks). The test trial ended as soon as the agent reached the Goal state. The model performance was quantified as in Eq. 1 and compared with the behavioral data: consistently with the data, we plotted the histogram of Δd_{ijk} for both populations and found that the model behaved qualitatively very similarly to the subjects (Fig. 8A, for a given set of parameters). Box-plots, showing Δd_{ijk} medians across observers and trials, confirmed this result also quantitatively (see inset in Fig. 8B).

Importantly, when EM were discarded, the model was unable to reproduce observers performance with the same accuracy; in the case of young subjects, while the algorithm performed optimally in 0 and A trials, its performance was severely suboptimal in B trials (left panel in Fig. 8C). The reason is fairly straightforward: at the end of the feeding procedure (i.e. at the end of the exploration phase), Q values of individual young subjects had, on average, converged to their optimal value. For the values associated with FDP, for example, $Q(U) > Q(L) > Q(D) > Q(R)$ (Fig. 9C). Accordingly, in the upcoming trial 0, the agent most likely chose option U while in FDP (since $\tau \simeq 0$), and successfully reached the Goal travelling trough the central corridor. In A and B trials, instead, U was negatively rewarded because of the block, and the agent chose the L option as the most rewarding one. The L choice successfully led to the Goal in A, but not in B trials. To correctly perform in B trials, standard RL algorithms should re-learn the new configuration of the environment, once blocks are introduced. The same reasoning holds true for old subjects; in this case, however, the variability of the algorithm performance was higher, given that Q values did not yet converge to their optimal values at the end of the exploration phase (right panel in Fig. 8C).

A similar suboptimal behaviour was found when EM were used to identify available paths to plan upon, but not to assign their corresponding rewards (Fig. 8D). In this

case, Eq. 8 become:

$$M(FDP, o_U) \leftarrow (s_U, r = 0) \tag{9}$$
$$M(FDP, o_R) \leftarrow (s_R, r = 0) \tag{10}$$
$$M(FDP, o_L) \leftarrow (s_L, r = 0) \tag{11}$$

and, as before, the environment should be re-learned for the algorithm to successfully perform on B trials.

Up to know we showed the model results for a single set of parameters. Next, we studied in which region of parameters these results hold. The relevant parameters were α, i.e. the learning rate, and n_{pl}, i.e. the number of planning steps. To compare the model performance to the data we considered the following quantity, for the young and the old population separately:

$$S = \langle \Delta d^M(\alpha, n_{pl}) - \Delta d^D \rangle_{ijk} \tag{12}$$

where k labelled the trial type ($k = 0, A, B$), j the number of repetitions of each k trial ($j = 1, 2, 3$), i the subject; the superscripts M and D indicated, respectively, model and data. In other words, we calculated the median of the deviation from the minimum path over subjects and trials for any given pair of parameters (α, n_{pl}) and compared it to the data; accordingly, the best parameters were those which minimized S.
We found that, young subjects data were best described by both large α and large n_{pl}; old subjects data instead were best described by smaller values of both parameters (blue regions in Fig. 9A). Consistently, plotting the model performance, $\langle \Delta d^M \rangle_{ijk}$, for the best parameters yielded results which were very close to the data for both populations (Fig. 9B). Note that, for the young population, the $0 < \alpha < 0.1$ region in the parameter space for which S appeared to be small is actually a local minimum (see Fig. 10 in Appendix).
We then looked at the dynamics of Q values in FDP obtained from the feeding procedure; we found that the parameters which best described the data were also those for which the Q values of the young population were optimal well before the end of the exploration phase, while those of the old population never reached optimality (Fig. 9C). This partly accounts for the flawless performance of the model in reproducing young data in the test trials and it provides an indication that young observers learned the structure of the environment already before the end of the exploration phase. Such a result was corroborated by verbal reports of young subjects at the end of the exploration phase, who spontaneously stated that 12min of exploration were unnecessary since they understood the maze structure already after the first couple of minutes.

The ensemble of these results quite naturally leads to the following interpretation: old subjects are slower learners (small α) and they plan their way through the maze less extensively (smaller n_{pl}). As a consequence, their behaviour is much closer to that of a model free algorithm rather than model based, as exemplified by their Q values dynamics (compare Fig. 9C. with Fig. 7C). It is worth noticing that the performance of the best observer among the elderly was best described by a much bigger n_{pl} than that describing the performance of the worst young subject. It follows that an optimal subject is one who plans his way through the maze more exhaustively (Fig. 10B in the Appendix).

Discussion

Despite the vast literature on spatial navigation, not much is known about how the brain adapts to unexpected changes in the environment (for a recent review see Spiers and Gilbert 2015); even less is known about how this ability is effected by a non-pathological aging process. Here, we relied on a RL framework to provide a mechanistic interpretation of how a detour task is solved. The model quite successfully reproduced the age related differences we found in the human ability to solve a Tolman detour task, in particular the impediment of older observers to grasp which alternative route would bring them faster to the Goal. The model supports two pivotal assertions on the role of eye movements in human planning strategies: i) the pattern of eye movements reflects the memory replay of experienced paths, which serves the planning of alternative routes; ii) the statistics of eye movements, i.e. the frequency with which observers look towards a given path, is a measure of the reward they expected when following it. The model's parameters which best fit the data indicated that older adults are slower learners but, indeed, they plan alternative paths through the maze like younger ones, although to a slightly less extent.

In the light of these results we put forward that global planning strategies do not radically change with age and that the impediment in solving a detour task found in older adults might rather be due to the building up of the wrong internal representation of the environment. In other words, older adults do plan, but over a wrong cognitive map.

Although it remains speculative, this hypothesis gets support from the maze sketched maps drawn by our older subjects, as well as from evidence that it takes a longer time for older adults to create an internal representation of the environment while navigating (Iaria et al. 2009). Consistently, the time we allocated for exploration in our paradigm might have been insufficient for older observers to build up a correct map of the maze. Moreover, older participants' pattern of eye movements points to a suboptimal strategy of acquisition of information about the environment during the exploration phase, which could also contribute to a defected map formation.

Crucially, we checked that the inability we found in older observers was not exclusively due to the lack of both proprioceptive and vestibular signals, which typically provide a reliable support to spatial navigation; when we asked observers to solve the Tolman detour task while actually walking through it, we found just about the same results obtained in the virtual setting (results not yet published).

On eye movements

An abundance of studies have shown that older adults are significantly impeded in spatial navigation tasks (for reviews see Moffat 2009; Lester et al. 2017). A bunch of these contributions measured performance differences between younger and older adults in tasks which supposedly relied on the ability to use and/or to form a cognitive map, finding, as we did, a detrimental effect of aging on spatial navigation performance (Newman and Kaszniak 2000; Moffat and Resnick 2002; Iaria et al. 2009). There is, however, to the best of our knowledge, no study which measured performance in a Tolman detour task in both younger and older observers and, more importantly, which analyzed the role of eye movements during a route re-planning strategy.

In the literature, eye movements have been strongly implicated in memory encoding

and retrieval; for example, limiting the number of fixations appears to be detrimental for both, when compared to free-viewing conditions (Henderson et al. 2005; Johansson and Johansson 2014). Moreover, altered fixation patterns in older adults have been proposed to be functionally linked to memory deficits (Chan et al. 2011; Hannula et al. 2007; Olsen et al. 2015; Ryan et al. 2000; Rondina et al. 2017; Shih et al. 2012; Voss et al. 2011). Consistently, we found that gaze direction distributions differed significantly between young and old observers during both the exploration and the probe trials (although we did not find any difference in the sheer number of fixations made; results not shown). In particular, the pattern of fixation of older observers exploring the maze, points to an inefficient encoding of stable information about the environment; likewise, the pattern of fixations recorded during the execution of the test trials is consistent with a suboptimal retrieval process, which, in older adults, leads up to a delayed decision about which path to choose (Fig. 5).

If the link between eye movements and memory is fairly well documented, much less is known about eye movements in the context of the re-planning of alternative routes. Two important pieces of evidence related eye movements and activity in the hippocampus, one of the brain structure most strongly implicated in the replay of past experiences as well as in the planning of future events (Buckner 2010; Foster and Wilson 2006; Gelbard-Sagiv et al. 2008; Javadi et al. 2017; Johnson and Redish 2007; Pastalkova et al. 2008; Pfeiffer and Foster 2013; Schacter et al. 2012; Wu and Foster 2014). First, during the encoding of novel stimuli, the number of fixations correlate with hippocampus activity, suggesting that visual sampling might be directly related to the formation of representations in the hippocampus (Liu et al. 2017). Second, the hippocampus guides where to look during memory retrieval (Hannula et al. 2007; Hannula and Greene 2012; Ryals et al. 2015; Ryan et al. 2000). These findings appear to be fully consistent with what we propose: eye movements subserve the retrieval of experienced paths from a mental representation of the environment, upon which future routes are planned. In other words, eye movements are the explicit manifestation of hippocampus-based computations implicated in planning. Our theory pushes the link between eye movements and planning even one step further: gaze directions associated with planning reflect the expectation of future rewards, meaning that we plan more often upon routes which we believe to be more rewarding and which, consequently, we deem more valuable. Furthermore, solving the detour task imply a re-evaluation of the value of the options available after the alteration of the environment (i.e. the introduction of the block in the maze); such an updating of reward assignment has been proposed to implicate hippocampal-striatal connectivity (Pennartz et al. 2011; Wimmer and Shohamy 2012).

Note that, here, we implied that planning consists in two concurrent processes i) the memory replay of experienced paths; ii) the re-evaluation, crucial for future decisions, of the benefits associated with any of these paths. Nonetheless, we acknowledge that planning does not necessarily rely on the recall of past experiences, e.g. in some circumstances, hippocampal cells are able to generate firing sequences representing paths never taken by the animal (Gupta et al. 2010; Ólafsdóttir et al. 2015). Moreover, it is important to stress that planning, in particular in the context of detours, is supported not only by the hippocampus but also by prefrontal cortices, although the relative contribution of the two is still unclear (Martinet et al. 2011; Spiers and Maguire 2006; Spiers and Gilbert 2015; Javadi et al. 2017). Furthermore, both ventromedial

prefrontal and striatal areas have been suggested to code for the value of the reward associated to mentally planned scenarios (Benoit et al. 2014; Lin et al. 2015).

On modelling

Reinforcement learning has been widely used as a learning rule in spatial navigation, not only for its efficiency, but also on the account of its biological plausibility. Midbrain dopaminergic neurons in the substantia nigra pars compacta (SNc) and ventral tegmental area (VTA) supply the striatum with dopamine, signalling a reward prediction error (Schultz et al. 1997), which represents the core of temporal-difference (TD) RL algorithms. The striatum itself is thought to be responsible for the computation of stimulus-response associations typical of MF versions of RL (Gläscher et al. 2010; Johnson et al. 2007; van Der Meer and Redish 2011). The more flexible learning rule implemented by MB algorithms, which relies on an explicit internal model of the environment, is instead carried out by the hippocampus (Gustafson and Daw 2011; Hasselmo 2005; Hirel et al. 2013; Martinet et al. 2011; Simon and Daw 2011).
Recently, the literature has converged on the idea that a combination of both MB and MF algorithms is most suitable to describe human navigation strategies, or, more generally human sequential decision making (Daw et al. 2005; Daw et al. 2011; Wunderlich et al. 2012; Gershman et al. 2014; Lee et al. 2014; Otto et al. 2014; Lee and Keramati 2017). In particular, the Dyna-Q algorithm (Sutton 1990) has been shown to be an optimal candidate, given that it relies by construction on a mixture of the two: the MB system, using an internal model of the environment, simulates offline state action transitions whose Q values are then updated through the MF system via a TD algorithm (Gershman et al. 2014).
Here, seeking computational efficiency, we implemented Dyna-Q in the framework of HRL (DQ-HRL). The advantage of partitioning the environment- as well as the actions therein- in larger chunks is well known computationally (Badre et al. 2010; Botvinick et al. 2009; Sutton and Barto 1998) and it has recently been corroborated, behaviourally, by the finding that the elaboration of plans in human observers is indeed hierarchical (Balaguer et al. 2016). We showed that DQ-HRL is faster in learning the shortest path to the Goal in a (block-free) Tolman maze, than both classical Dyna-Q and the model free version of HRL (MF-HRL; Fig. 7). This is mainly because classical algorithms need to learn the value of each state and action within each arm of the maze, which would entail wasting time in going back and forth through sub-portions of a selected arm. To the contrary, an algorithm like HRL would learn the value of an arm as a whole, resulting in much faster learning.
However, although extremely efficient, even HRL would need to re-learn the new structure of the environment once the blocks are introduced and, consequently, it would always perform poorly with respect to a young observer, who is capable of taking the optimal detour without further training. In other words, in any RL algorithm the agent would need to visit several time the state immediately close to the block to learn to attribute the appropriate value to it and to be able to choose a suitable alternative (see for example Russek et al. 2017). However, none of our subjects during the test trials needed to do so, i.e. none of our subjects ever walked straight to the block: looking at the obstacle from a distance was enough to grasp that the central corridor was occluded and that one of the other available paths had to be chosen instead. Inspired by this observation, we gave the agent in the DQ-HRL algorithm the ability to plan over the

states our observers only looked at, ending up with a RL model which, for the first time, can learn the Tolman detour task in one-shot.
Upon fitting the subjects' trajectories and eye-movements, our model predicts that older adults are not only slower learners, but also less efficient planners compared to younger adults. Plausibly, ageing has an effect on the quality of the internal representation of the environment on which planning is based. Future work might focus on pushing the eye-movements analysis to a higher level of sophistication, e.g. to understand in which respect a cognitive map might result inaccurate and whether ageing affects specifically the binding of the different structural elements of the environment together or rather other aspects of the map formation.

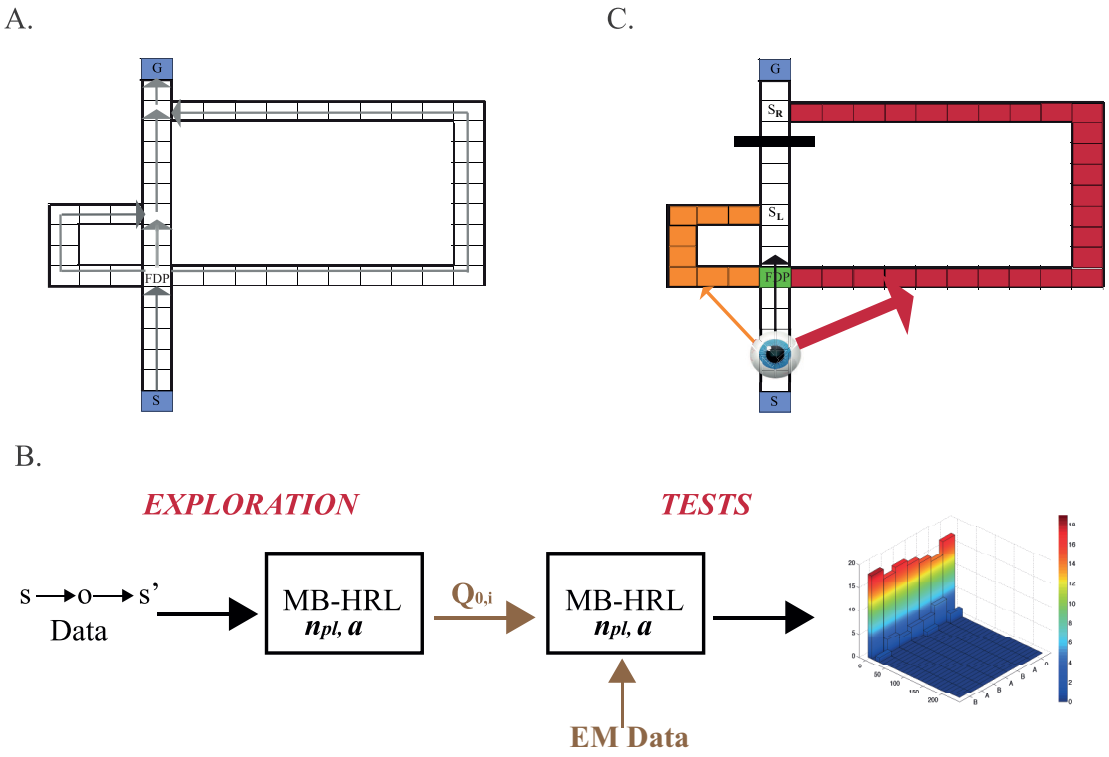


Figure 6: **A.** For modelling purposes, the Tolman maze is decomposed in a set of 52 states (black squares in the cartoon) and 6 non-overlapping *options* (grey arrows). Each option is bidirectional, although, for clarity, only one direction is depicted in the figure. The length k of an option is given by the number of underlying states it covers. For example, the first option initiates at the starting point of the maze (S) and terminates at the first decision point (FDP), hence $k = 6$. **B.** The model was first "fed" with the individual subjects trajectories measured during the exploration phase, and produced the Q values of each individual subject as output. Those values, together with individual subjects eye-movements, were used to probe the model with the sequence of behavioural test trials 0-A-B-A-B-A-B. In both the exploration and the test phase the relevant parameters were the number of planning steps n_{pl} and the learning rate α. **C.** Subjects eye-movements were used in the model to modulate the reward associated with the options available at FDP during the test trials. In the cartoon a B trial is depicted. If a subject, navigating from S to FDP, fixated 20% of the time towards the right, the reward associated to the transition $FDP \rightarrow s_R$ was set to $r = 0.2$. Likewise, if the proportion of fixation to the left was 10%, the reward for the transition $FDP \rightarrow s_L$ was set to $r = 0.1$.

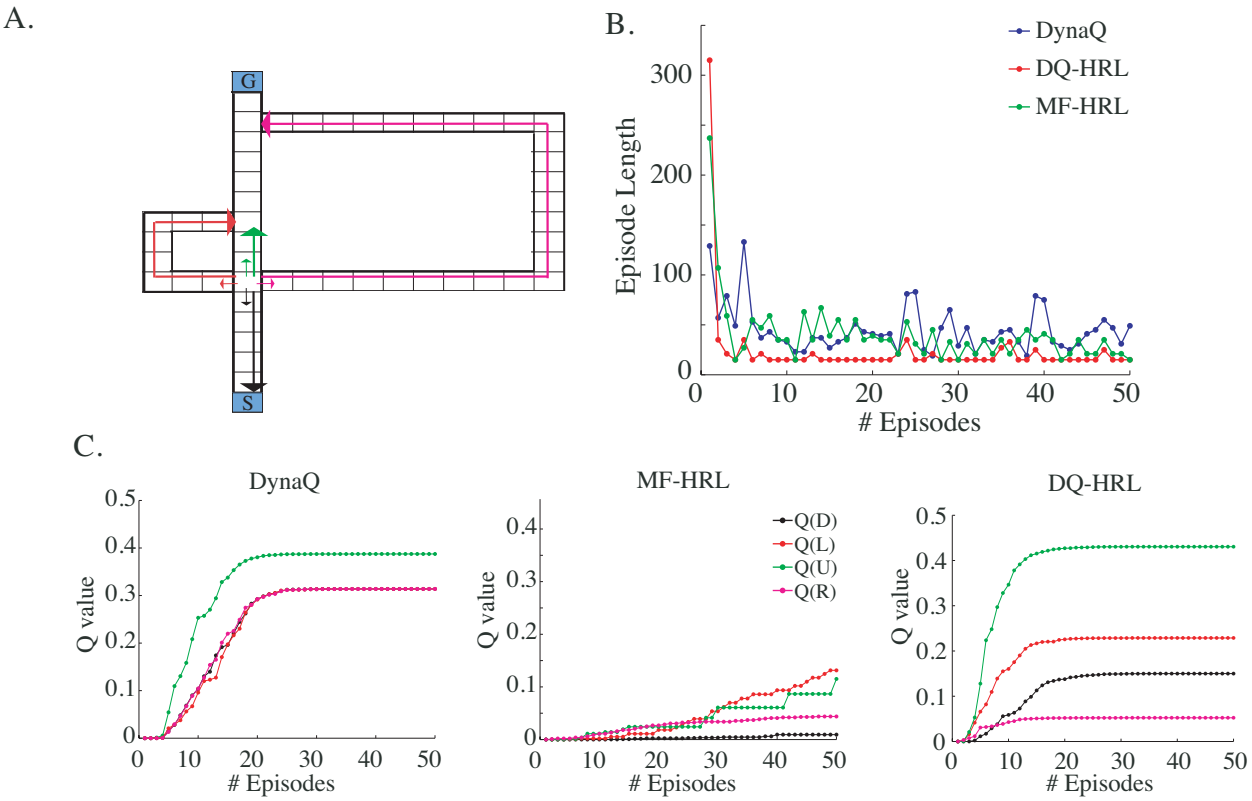


Figure 7: **A.** Cartoon of a subset of available *options*, used in HRL algorithms, versus *primitive actions*, used in standard RL algorithms. The former are a spatially extended version of the latter. **B.** Performance of model based HRL, model-free HRL and model-based standard RL (i.e. Dyna-Q) measured as in Eq. 1 for the Tolman maze shown in A. **C** Dynamics of Q values in FDP for the three algorithms shown in B. The optimal algorithm should converge to $Q(U) > Q(L) > Q(D) > Q(R)$, as $DQ - HRL$ does. Parameters: $\alpha = 0.1$; $\gamma = 0.9$; $\tau = 0.1$; $n_{pl} = 10$ for model based; $n_{pl} = 0$ for model free.

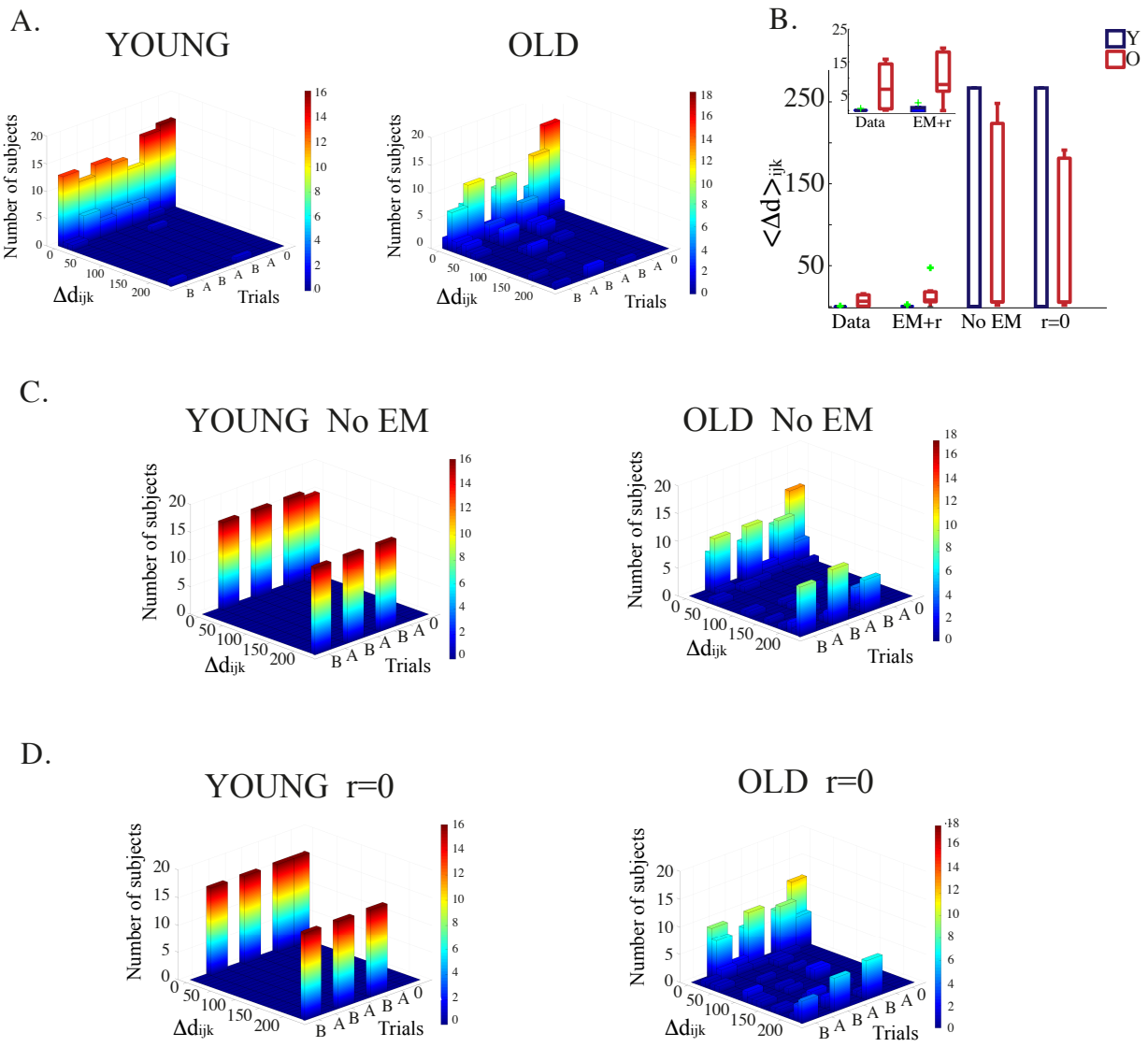


Figure 8: **A.** Model performance measured, analogously to the data, as the distance travelled by individual observers in each test trial. To obtain these results, individual exploration data as well as individual EM have been injected in the *HRL* algorithm. **B.** Median across observers and trials of the travelled distance, calculated for the data, for the model using EM statistics and associated reward modulation, for the model without EM and for the model in which EM do not entail a reward modulation. Disregarding EM and/or reward modulation is significantly detrimental for the model performance. **C.** Same as in A., but for the model which does not use EM. **D.** Same as in A., but for the model which use EM only to select paths to plan upon, without associating any reward to them. Parameters: young population: $\alpha = 0.21$; $\gamma = 0.9$; $\tau = 0.01$; $n_{pl} = 29$; old population: $\alpha = 0.01$; $\gamma = 0.9$; $\tau = 0.01$; $n_{pl} = 17$;

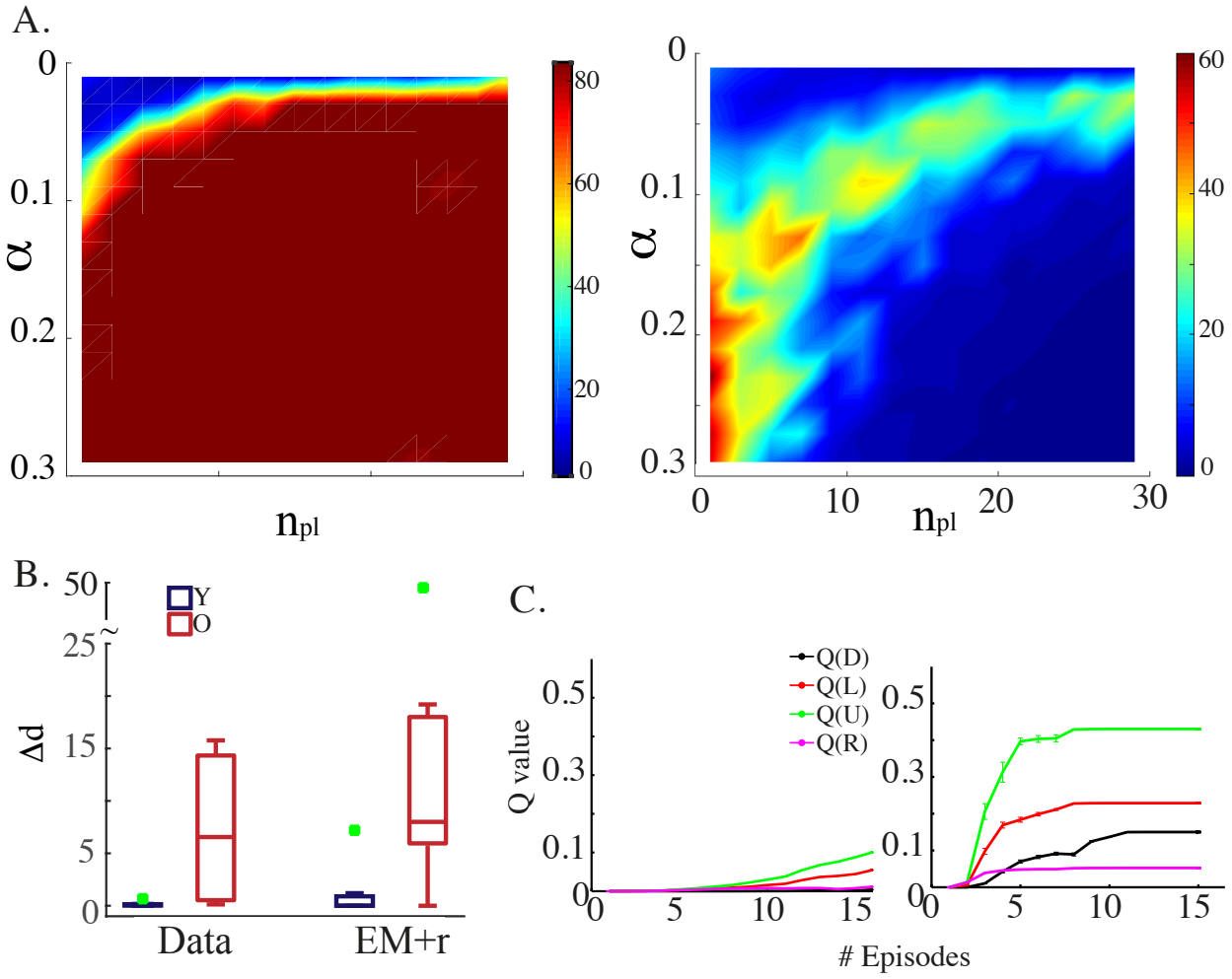


Figure 9: **Sensitivity anlaysis A.** Difference in performance between data and model, calculated as in Eq. 12, as a function of the learning rate α and the number of planning steps n_{pl}. The best parameters are: young population: $\alpha^* = 0.19$; $n^*_{pl} = 25$; old population: $\alpha^* = 0.01$; $n^*_{pl} = 17$. **B.** Median across observers and trials of the travelled distance, calculated for the data and for the model with $\alpha = \alpha^*$ and $n_{pl} = n^*_{pl}$. Average Q values dynamics in FDP as a result of the feeding procedure in which we set $\alpha = \alpha^*$ and $n_{pl} = n^*_{pl}$. Error bars represent s.e.m. For the exploration data used in the feeding procedure, the average number of episodes across observers was 10.27+2.53, while the average number of state transitions (i.e. the average number of options taken) was 64.16+12.21.

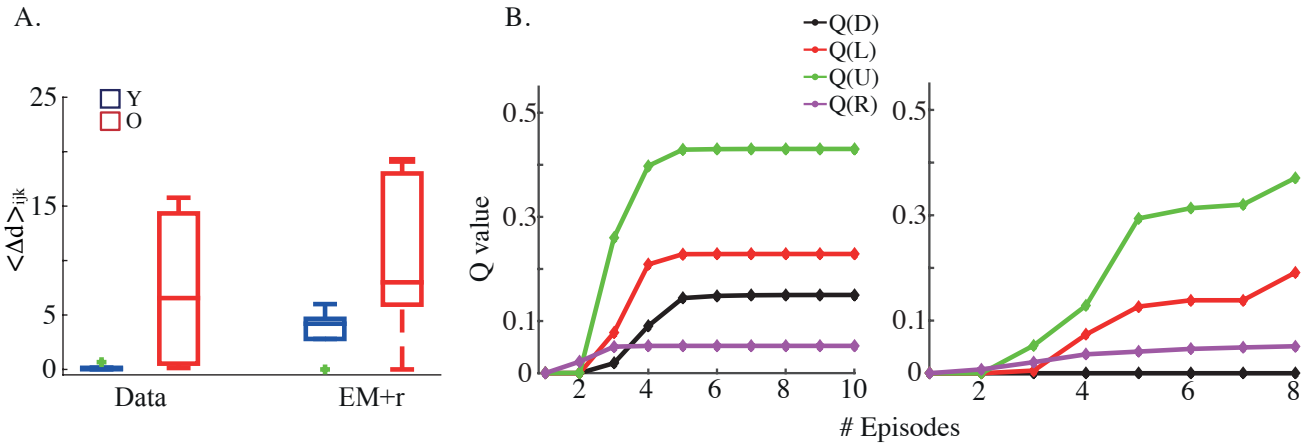


Figure 10: **A.** Median across observers and trials of the travelled distance, calculated for the data and for the model with α and n_{pl} set at the local minimum for the young population, which corresponds to $\alpha = 0.01$; $n_{pl} = 15$; **B.** Left: Average Q values dynamics in FDP for the best old subject, for whom $\alpha = 0.21$ and $n_{pl} = 21$. Right: Average Q values dynamics in FDP for the worst young subject, for whom $\alpha = 0.21$ and $n_{pl} = 3$.

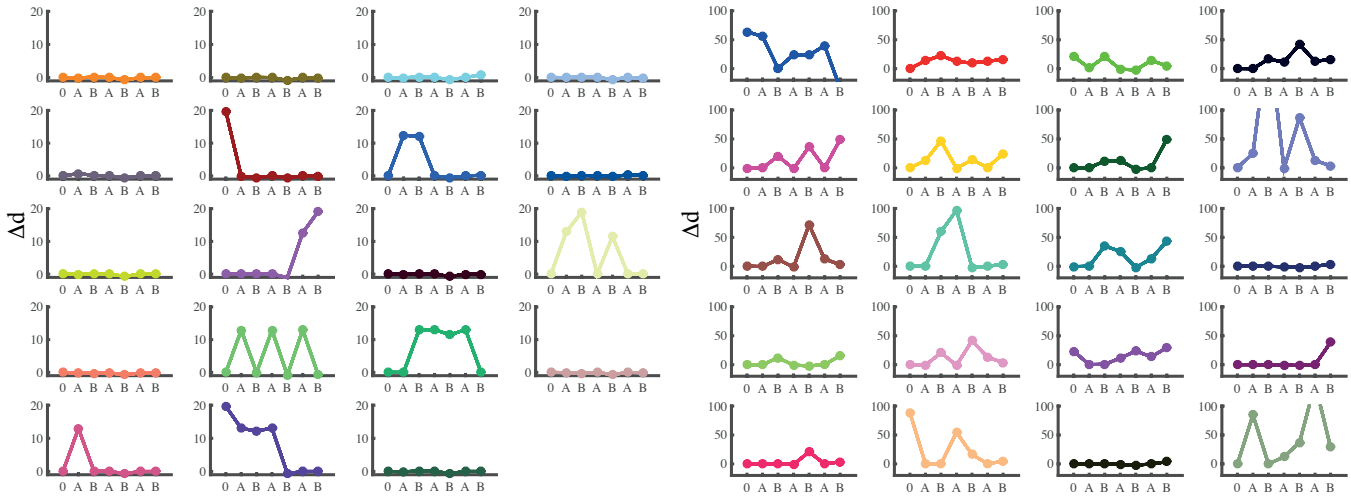
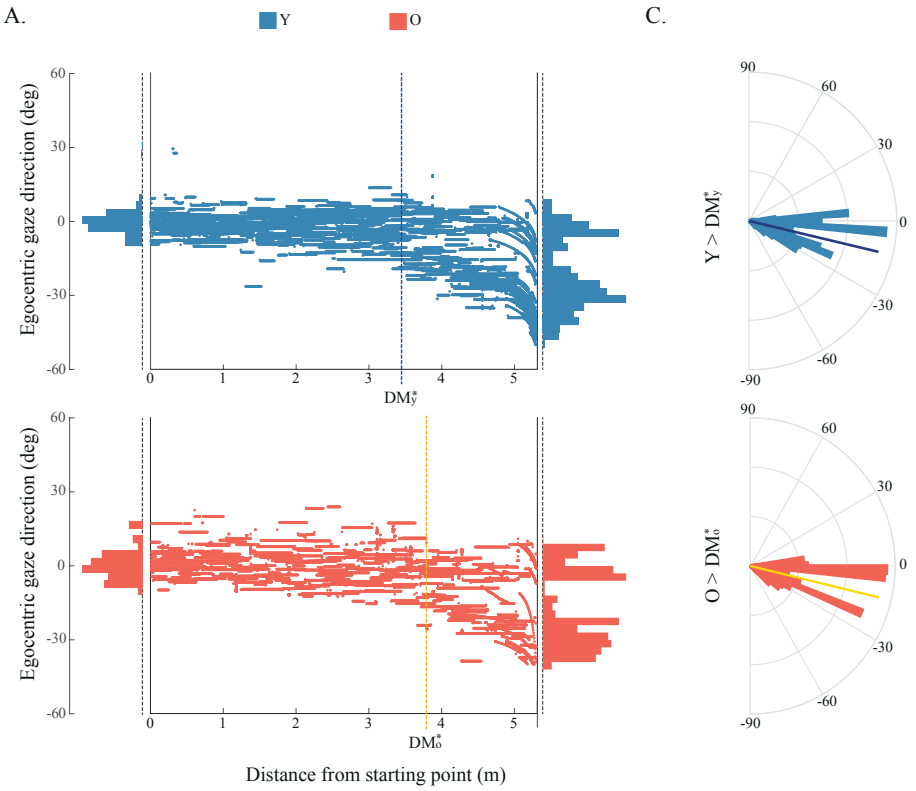


Figure 11: **A. B.** .

Sketched map evaluation criteria	**Score**
A main alley (north-south axis) exists	+1
The Goal location is facing the Start location	+1
The southernmost intersection with the main alley is cross-shaped (intersection 1/point 1)	+2
An intersection located to the left of the main alley exists (intersection 2/point 2)	+2
An intersection located to the left of the main alley exists (intersection 3/point 3). Intersection 2 and 3 can be at the same location forming a cross-shaped intersection	+2
Intersection 2 is located south of intersection 3	+2.5
Points 1 and 2 are connected through a path which is not the main alley	+2
Points 1 and 3 are connected through a path which is not the main alley	+2
Points 1 and 2 are connected via 3 segments forming a 90 deg with each other	+1
Points 1 and 3 are connected with 3 segments forming a 90 deg with each other	+1
The path connecting point 1 and 2 is shorter than the path connecting point 1 and 3	+2.5
No-entry signs exist between point 1 and point 2 and between point 2 and point 3	+0.5 each
Extra alleys are drawn	-0.5 each
Alleys leading nowhere are drawn (not connected to the maze)	-0.5 each
Extra intersections are drawn	-0.5 each
Extra "no-entry" signs are drawn	-0.5 each



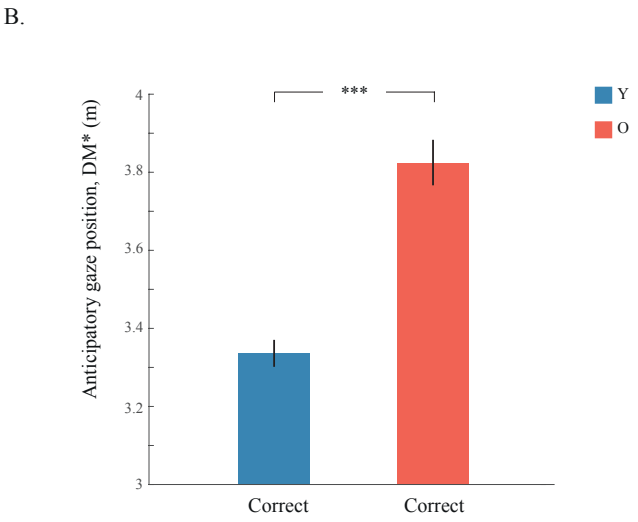


Figure S??

Figure 12: Same as Fig. 5 but for 'Correct' Y and 'Correct' O datasets.

References

Antonova, E., D. Parslow, M. Brammer, G. Dawson, S. Jackson, and R. Morris (2009). Age-related neural activity during allocentric spatial memory. *Memory 17*(2), 125–143.

Badre, D., A. S. Kayser, and M. D'Esposito (2010). Frontal cortex and the discovery of abstract action rules. *Neuron 66*(2), 315–326.

Balaguer, J., H. Spiers, D. Hassabis, and C. Summerfield (2016). Neural mechanisms of hierarchical planning in a virtual subway network. *Neuron 90*(4), 893–903.

Barrash, J. (1994). Age-related decline in route learning ability. *Developmental Neuropsychology 10*(3), 189–201.

Benoit, R. G., K. K. Szpunar, and D. L. Schacter (2014). Ventromedial prefrontal cortex supports affective future simulation by integrating distributed knowledge. *Proceedings of the National Academy of Sciences 111*(46), 16550–16555.

Bernardin, D., H. Kadone, D. Bennequin, T. Sugar, M. Zaoui, and A. Berthoz (2012). Gaze anticipation during human locomotion. *Experimental brain research 223*(1), 65–78.

Berns, G. S., S. M. McClure, G. Pagnoni, and P. R. Montague (2001). Predictability modulates human brain response to reward. *The journal of neuroscience 21*(8), 2793–2798.

Billinghurst, M. and S. Weghorst (1995). The use of sketch maps to measure cognitive maps of virtual environments. In *Virtual Reality Annual International Symposium, 1995. Proceedings.*, pp. 40–47. IEEE.

Bohbot, V. D., S. McKenzie, K. Konishi, C. Fouquet, V. Kurdi, R. Schachar, M. Boivin, and P. Robaey (2012). Virtual navigation strategies from childhood to senescence: evidence for changes across the life span. *Frontiers in aging neuroscience 4*.

Botvinick, M. M., Y. Niv, and A. C. Barto (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition 113*(3), 262–280.

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual review of psychology 61*, 27–48.

Burns, P. C. (1999). Navigation and the mobility of older drivers. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 54*(1), S49–S55.

Chan, J. P., D. Kamino, M. A. Binns, and J. D. Ryan (2011). Can changes in eye movement scanning alter the age-related deficit in recognition memory? *Frontiers in psychology 2*.

Chersi, F. and N. Burgess (2015). The cognitive architecture of spatial navigation: Hippocampal and striatal contributions. *Neuron 88*(1), 64–77.

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

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

Driscoll, I., D. A. Hamilton, R. A. Yeo, W. M. Brooks, and R. J. Sutherland (2005). Virtual navigation in humans: the impact of age, sex, and hormones on place learning. *Hormones and behavior 47*(3), 326–335.

Foster, D. J. and M. A. Wilson (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature 440*(7084), 680–683.

Gazova, I., J. Laczó, E. Rubinova, I. Mokrisova, E. Hyncicova, R. Andel, M. Vyhnalek, K. Sheardova, E. J. Coulson, and J. Hort (2013). Spatial navigation in young versus older adults. *Frontiers in aging neuroscience 5*.

Gelbard-Sagiv, H., R. Mukamel, M. Harel, R. Malach, and I. Fried (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science 322*(5898), 96–101.

Gershman, S. J., A. B. Markman, and A. R. Otto (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General 143*(1), 182.

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

Gupta, A. S., M. A. van der Meer, D. S. Touretzky, and A. D. Redish (2010). Hippocampal replay is not a simple function of experience. *Neuron 65*(5), 695–705.

Gustafson, N. J. and N. D. Daw (2011). Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLoS computational biology 7*(10), e1002235.

Hannula, D. E. and A. J. Greene (2012). The hippocampus reevaluated in unconscious learning and memory: at a tipping point? *Frontiers in human neuroscience 6*.

Hannula, D. E., J. D. Ryan, D. Tranel, and N. J. Cohen (2007). Rapid onset relational memory effects are evident in eye movement behavior, but not in hippocampal amnesia. *Journal of cognitive neuroscience 19*(10), 1690–1705.

Harris, M. A. and T. Wolbers (2014, May). How age-related strategy switching deficits affect wayfinding in complex environments. *Neurobiology of Aging 35*(5), 1095–1102.

Hasselmo, M. E. (2005). A model of prefrontal cortical mechanisms for goal-directed behavior. *Journal of cognitive neuroscience 17*(7), 1115–1129.

Head, D. and M. Isom (2010). Age effects on wayfinding and route learning skills. *Behavioural brain research 209*(1), 49–58.

Henderson, J. M., C. C. Williams, and R. J. Falk (2005). Eye movements are functional during face learning. *Memory & cognition 33*(1), 98–106.

Hirel, J., P. Gaussier, M. Quoy, J.-P. Banquet, E. Save, and B. Poucet (2013). The hippocampo-cortical loop: spatio-temporal learning and goal-oriented planning in navigation. *Neural Networks 43*, 8–21.

Iaria, G., L. Palermo, G. Committeri, and J. J. Barton (2009). Age differences in the formation and use of cognitive maps. *Behavioural brain research 196*(2), 187–191.

Jansen, P., A. Schmelter, and M. Heil (2010). Spatial knowledge acquisition in younger and elderly adults: A study in a virtual environment. *Experimental Psychology 57*(1), 54.

Javadi, A.-H., B. Emo, L. R. Howard, F. E. Zisch, Y. Yu, R. Knight, J. P. Silva, and H. J. Spiers (2017). Hippocampal and prefrontal processing of network topology to simulate the future. *Nature Communications 8*, 14652.

Johansson, R. and M. Johansson (2014). Look here, eye movements play a functional role in memory retrieval. *Psychological Science 25*(1), 236–242.

Johnson, A. and A. D. Redish (2007). Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience 27*(45), 12176–12189.

Johnson, A., M. A. van der Meer, and A. D. Redish (2007). Integrating hippocampus and striatum in decision-making. *Current opinion in neurobiology 17*(6), 692–697.

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

Lee, J. and M. Keramati (2017). Flexibility to contingency changes distinguishes habitual and goal-directed strategies in humans. *bioRxiv*, 107078.

Lee, S. W., S. Shimojo, and J. P. ODoherty (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron 81*(3), 687–699.

Lester, A. W., S. D. Moffat, J. M. Wiener, C. A. Barnes, and T. Wolbers (2017). The aging navigational system. *Neuron 95*(5), 1019–1035.

Lin, W.-J., A. J. Horner, J. A. Bisby, and N. Burgess (2015). Medial prefrontal cortex: adding value to imagined scenarios. *Journal of cognitive neuroscience*.

Liu, Z.-X., K. Shen, R. K. Olsen, and J. D. Ryan (2017). Visual sampling predicts hippocampal activity. *Journal of Neuroscience 37*(3), 599–609.

Madl, T., K. Chen, D. Montaldi, and R. Trappl (2015). Computational cognitive models of spatial memory in navigation space: A review. *Neural Networks 65*, 18–43.

Martinet, L.-E., D. Sheynikhovich, K. Benchenane, and A. Arleo (2011). Spatial learning and action planning in a prefrontal cortical network model. *PLoS computational biology 7*(5), e1002045.

Moffat, S. D. (2009). Aging and spatial navigation: what do we know and where do we go? *Neuropsychology review 19*(4), 478–489.

Moffat, S. D., W. Elkins, and S. M. Resnick (2006). Age differences in the neural systems supporting human allocentric spatial navigation. *Neurobiology of aging 27*(7), 965–972.

Moffat, S. D. and S. M. Resnick (2002). Effects of age on virtual environment place navigation and allocentric cognitive mapping. *Behavioral neuroscience 116*(5), 851.

Newman, M. C. and A. W. Kaszniak (2000, June). Spatial Memory and Aging: Performance on a Human Analog of the Morris Water Maze. *Aging, Neuropsychology, and Cognition 7*(2), 86–93.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current opinion in neurobiology 14*(6), 769–776.

Ólafsdóttir, H. F., C. Barry, A. B. Saleem, D. Hassabis, and H. J. Spiers (2015). Hippocampal place cells construct reward related sequences through unexplored space. *Elife 4*, e06063.

Olsen, R. K., Y. Lee, J. Kube, R. S. Rosenbaum, C. L. Grady, M. Moscovitch, and J. D. Ryan (2015). The role of relational binding in item memory: evidence from face recognition in a case of developmental amnesia. *Journal of Neuroscience 35*(13), 5342–5350.

Otto, A. R., A. Skatova, S. Madlon-Kay, and N. D. Daw (2014). Cognitive control predicts use of model-based reinforcement learning. *Journal of cognitive neuroscience*.

Pagnoni, G., C. F. Zink, P. R. Montague, and G. S. Berns (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature neuroscience 5*(2), 97–98.

Pastalkova, E., V. Itskov, A. Amarasingham, and G. Buzsaki (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science 321*, 1322–1327.

Pennartz, C., R. Ito, P. Verschure, F. Battaglia, and T. Robbins (2011). The hippocampal–striatal axis in learning, prediction and goal-directed behavior. *Trends in neurosciences 34*(10), 548–559.

Pfeiffer, B. E. and D. J. Foster (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature 497*(7447), 74–79.

Rodgers, M. K., J. A. Sindone, and S. D. Moffat (2012). Effects of age on navigation strategy. *Neurobiology of aging 33*(1), 202–e15.

Rondina, R., K. Curtiss, J. A. Meltzer, M. D. Barense, and J. D. Ryan (2017). The organisation of spatial and temporal relations in memory. *Memory 25*(4), 436–449.

Russek, E. M., I. Momennejad, M. M. Botvinick, S. J. Gershman, and N. D. Daw (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *bioRxiv*, 083857.

Ryals, A. J., J. X. Wang, K. L. Polnaszek, and J. L. Voss (2015). Hippocampal contribution to implicit configuration memory expressed via eye movements during scene exploration. *Hippocampus 25*(9), 1028–1041.

Ryan, J. D., R. R. Althoff, S. Whitlow, and N. J. Cohen (2000). Amnesia is a deficit in relational memory. *Psychological Science 11*(6), 454–461.

Schacter, D. L., D. R. Addis, D. Hassabis, V. C. Martin, R. N. Spreng, and K. K. Szpunar (2012). The future of memory: remembering, imagining, and the brain. *Neuron 76*(4), 677–694.

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

Shih, S.-I., K. L. Meadmore, and S. P. Liversedge (2012). Aging, eye movements, and object-location memory. *PLoS One 7*(3), e33485.

Simon, D. A. and N. D. Daw (2011). Neural correlates of forward planning in a spatial decision task in humans. *The Journal of Neuroscience 31*(14), 5526–5539.

Spiers, H. J. and S. J. Gilbert (2015). Solving the detour problem in navigation: a model of prefrontal and hippocampal interactions. *Frontiers in human neuroscience 9*.

Spiers, H. J. and E. A. Maguire (2006). Thoughts, behaviour, and brain dynamics during navigation in the real world. *Neuroimage 31*(4), 1826–1840.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the seventh international conference on machine learning*, pp. 216–224.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement learning: An introduction*, Volume 1. MIT press Cambridge.

Tartaglia, E. M., A. Clarke, and M. H. Herzog (2017). Where to go next? A paradigm for testing human sequential decision making. *Frontiers in Neuroscience*, Under Revision.

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

van Der Meer, M. A. and A. D. Redish (2011). Ventral striatum: a critical look at models of learning and evaluation. *Current opinion in neurobiology 21*(3), 387–392.

Voss, J. L., D. E. Warren, B. D. Gonsalves, K. D. Federmeier, D. Tranel, and N. J. Cohen (2011). Spontaneous revisitation during visual exploration as a link among strategic behavior, learning, and the hippocampus. *Proceedings of the National Academy of Sciences 108*(31), E402–E409.

Wilkniss, S. M., M. G. Jones, D. L. Korol, P. E. Gold, and C. A. Manning (1997). Age-related differences in an ecologically based study of route learning. *Psychology and aging 12*(2), 372.

Wimmer, G. E. and D. Shohamy (2012). Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science 338*(6104), 270–273.

Wu, X. and D. J. Foster (2014). Hippocampal replay captures the unique topological structure of a novel environment. *Journal of Neuroscience 34*(19), 6459–6469.

Wunderlich, K., P. Smittenaar, and R. J. Dolan (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron 75*(3), 418–424.