scientific reports

Check for updates

OPEN

A comparison of reinforcement learning models of human spatial navigation

Qiliang He[1]✉, Jancy Ling Liu[2], Lou Eschapasse[1], Elizabeth H. Beveridge[1] & Thackery I. Brown[1]✉

Reinforcement learning (RL) models have been influential in characterizing human learning and decision making, but few studies apply them to characterizing human spatial navigation and even fewer systematically compare RL models under different navigation requirements. Because RL can characterize one's learning strategies quantitatively and in a continuous manner, and one's consistency of using such strategies, it can provide a novel and important perspective for understanding the marked individual differences in human navigation and disentangle navigation strategies from navigation performance. One-hundred and fourteen participants completed wayfinding tasks in a virtual environment where different phases manipulated navigation requirements. We compared performance of five RL models (3 model-free, 1 model-based and 1 "hybrid") at fitting navigation behaviors in different phases. Supporting implications from prior literature, the hybrid model provided the best fit regardless of navigation requirements, suggesting the majority of participants rely on a blend of model-free (route-following) and model-based (cognitive mapping) learning in such navigation scenarios. Furthermore, consistent with a key prediction, there was a correlation in the hybrid model between the weight on model-based learning (i.e., navigation strategy) and the navigator's exploration vs. exploitation tendency (i.e., consistency of using such navigation strategy), which was modulated by navigation task requirements. Together, we not only show how computational findings from RL align with the spatial navigation literature, but also reveal how the relationship between navigation strategy and a person's consistency using such strategies changes as navigation requirements change.

Reinforcement learning (RL) has made tremendous progress in the past two decades in computer science, psychology and neuroscience[1–5]. RL describes a learning mechanism in which behaviors are shaped through approaching rewards and avoiding punishments, which has a long history dating back to the nineteenth century[6]. In psychology, RL studies show that RL can be hierarchically structured[7] and it interacts with other cognitive functions such as working memory in decision-making[8–10]. In neuroscience, RL studies have sought to reveal the neural substrates representing RL parameters (e.g., prediction error in the RL model), which include the striatum[11,12], the ventromedial prefrontal cortex[13], hippocampus[14] and the firing behavior of dopamine neurons[15]. Most of the paradigms used in human RL studies focus on decision-making in 2D environments, so it remains understudied whether these findings can be generalized to scenarios important for survival like spatial navigation, which takes place in 3D environments. Moreover, modeling human spatial navigation behaviors via RL models could provide additional insight into the underlying cognitive mechanisms for a given navigational route compared to traditional methods (e.g., computing a route's length) by using a person's own navigational history to quantify that individual's navigation strategies and their consistency of using such strategies (discussed more below). The current study aims to address these gaps and opportunities in the literature.

To maximize rewards while minimizing effort/cost, we learn the associations between behaviors and rewards mainly through two types of RL: model-free—which fosters rigid repetitions of previously rewarded actions, or model-based—which fosters a mental model of the environment or task structure to flexibly select goal-directed actions[11,12,16,17]. In spatial navigation, model-free learning corresponds to response learning, or merely learning the landmark-action associations (e.g., turn right at the second traffic light) without learning the overall layout of the environment. Model-based learning, on the other hand, corresponds to place learning or learning the configuration of the environment[12,18,19]. From the RL literature, there is evidence showing that model-free and

[1]School of Psychology, Georgia Institute of Technology, Atlanta, USA. [2]School of Economics, Georgia Institute of Technology, Atlanta, USA. ✉email: duncan.heqiliang@gmail.com; thackery.brown@psych.gatech.edu

model-based learning operate and compete in parallel[9,11,20,21]. From the human spatial cognition literature, there is a plethora of evidence showing there are substantial individual differences in human spatial navigation, including wayfinding performance[19,22–28] and navigation strategy[29–34]. Such individual differences are found to be correlated with structural and functional differences of the brain[34–39], working memory capacity[23,27,40], gender[29,41] and task instructions[31,33]. Taking the findings from RL and spatial cognition together, it is reasonable to hypothesize that whereas the navigation behaviors of some people are best fit by either the model-free or the model-based learning, the majority of them should be best fit by a hybrid model, which is a combination or more-continuous balance of model-free and model-based learning.

Surprisingly, there were very few studies lending support to this idea that such a hybrid RL model provides the best fit to characterize human spatial navigation. Using a spatial navigation task where the layout of the environment changed continuously, Simon and Daw[12] found that the model-based learning fit the data better than the model-free, but whether a hybrid model provides a better fit than the model-based model remains unknown. Using several spatial navigation tasks, Anggraini et al.[18] compared the neural substrates which tracked model-free, model-based and hybrid parameters, but which model provides the best fit to the navigation data themselves remains unknown. We believe that comparing the model-free, model-based and the hybrid models is crucial for efforts to investigate how the results from computational modeling, such as RL, fit the well-established individual differences and strategy-preference findings in the spatial cognition literature.

If the results from RL are in line with the findings from the literature, then we can harness the parameters generated from RL models to better understand human spatial navigation with much more confidence. As noted above, one unique contribution of applying RL to understanding human spatial navigation is that RL models can reveal one's navigation strategy quantitatively and as a function of one's prior experiences. In recent literature, we and others have used the dual-solution task[30] to quantify individuals' navigation strategies towards following a familiar route vs. taking a shortcut. In this approach, participants first follow a pre-defined route to a destination several times, and then they can navigate to the destination freely. Using this task, individual navigation strategy is quantified via the solution index, which is equal to the number of trials in which a novel shortcut is taken divided by the number of successful trials in which either a shortcut or the learned route is taken. Studies using this task show marked individual differences in navigation tendency, such that some individuals primarily use shortcuts, some primarily use familiar routes, and many fall between these extremes[30,31,34]. However, classifying the navigation strategy of an episode into either familiar route following or shortcut taking in this manner may simplify many complex navigation behaviors observed in our daily life. It is often possible that one first follows the first few sections of the learned route and then takes a shortcut from there; similarly, the cognitive demands of a "shortcut" may depend heavily on how much of the route sequence has been previously traversed (i.e., to what degree is the route a novel construct). The weight parameter of a RL hybrid model can reveal a navigator's reliance on cognitive map vs. route-following on a trial-by-trial basis, which provides a finer grained characterization of one's navigation strategy in an objective and continuous manner, based on their prior navigation history. Furthermore, RL models can also estimate the consistency of using a navigation strategy across different navigation episodes. These two parameters from RL provide important insight into how one adapts their navigation strategies under different navigation requirements, which is a great complement to the solution index of the dual-solution task.

To select the appropriate RL algorithms for the current study, we used temporal difference (TD) learning[9,11,12,18,21], which is one of the most commonly used model-free algorithms in the RL literature. TD learning assumes that agents learn the future reward value following an action, and adjusts predictions continuously before obtaining the reward[5]. In the current study, we compared three types of TD models: TD(0), TD(λ) and TD(1). The major difference between TD(λ) and TD(0) is that TD(λ) adds an eligibility trace (see "Methods"), which adds the assumption that all the values of the visited locations are updated over time and the amount of updating depends on the visitation frequency. TD(1) is a special case of TD(λ) such that once a location is visited, the updated value at a location would never diminish over time even if it is not visited again. In other words, TD(1) assumes that there is no forgetting of the importance of the visited locations and therefore there is no need for a navigator whose cognition resembles this framework to revisit them to retain their relevance in wayfinding. We target these TD models because they differ from each other on how the importance of the visited locations changes over time—essentially, they make different assumptions about memory updating in spatial navigation. The topic of memory updating in spatial navigation has been studied extensively in experimental approaches[33,42–46], but rarely through this computational lens. Note that these three TD algorithms predict that participants select different paths to reach destination based on their navigation history, but do not necessarily predict different navigation performance [e.g., one who favors TD(1) does not necessarily have to have better performance than the one who favors TD(0)].

In addition to model-free learning, we constructed a model-based model for spatial navigation. People who completely rely on a model-based system are assumed to have a perfect cognitive map[47,48], and therefore one model-based model is sufficient. Because human performance relying purely on idealized cognitive maps may be implausible (at least for the vast majority of people) in many navigational scenarios, a hybrid model was constructed to reflect a heterogeneity or balance within an individual between map-like and more experience-bound knowledge. This hybrid model was developed by combing the best performing TD model with the model-based model, and adds a free weight parameter (ω) to capture the individual's relative reliance on the model-based learning. As mentioned earlier, we hypothesized that the hybrid model was the best performing model in fitting human spatial navigation data.

The results from the model comparisons can inform us how well RL models fit the human spatial navigation literature, and as mentioned above, there are at least two parameters (navigation strategies and the consistency of using such strategies) generated from the RL models which could let us gain additional insight on human spatial navigation compared to traditional methods. To this end, we created navigation tasks with different requirements

and investigated how these different navigational scenarios modulated navigation strategies and the consistency of using such strategies, which reflected how humans adapted to the ever-changing environment[12] but has rarely been examined in the literature. Specifically, our design considers the fact that not every goal-directed navigation problem is best approached in the same way, and this creates a dynamic in which individual differences can also be understood in terms of how people shift their learning/behavioral model under different demands. In the current study, participants first found different objects in a virtual environment from a fixed starting location (the Fixed phase), and then found the same objects in the same environment but from various random locations (the Random phase). We were interested in how the relative model-based weight (ω) and the exploration–exploitation parameter (θ) changed as a function of learning experience and these different task requirements. In our hybrid model, ω represented the mean of an individual's navigation strategy across a number of navigation trials, and θ represents the *consistency* of using this strategy in these trials (i.e., the degree to which an individual persists with a particular way of approaching the tasks in the face of feedback and changing demands).

Examining ω and θ separately and jointly would shed important light on how humans adapted to navigation scenarios of different requirements. Based on our manipulation of the navigation requirements, we hypothesized that (1) ω, the reliance on model-based system or the cognitive map, would increase from the Fixed to the Random phases due to increasing familiarity of the environment[19,22,23,49] and the demands of the Random phase encouraging greater reliance on map-like knowledge. (2) Participants would be more exploratory or deviate from their default strategy more in the Random phase due to the randomness and uncertainty introduced[50]. Therefore, we hypothesized that θ would increase from the Fixed to the Random phases. (3) The correlations between ω and θ would be different in the Fixed phase from the Random phase—in the Fixed phase, where the starting location was always the same, there was no need to vary navigation strategy from trial to trial. In the Random phase, on the other hand, a more efficient strategy would be to rely on the model-free system when starting from a familiar location but to rely on the model-based system when starting from an unfamiliar location (thus favoring variation of navigation strategy). In other words, we theorized that better navigators would use one strategy more consistently in deterministic navigational scenarios, whereas they would vary their strategy more often in probabilistic navigational scenarios. From this theoretical perspective, we hypothesized that the correlation between ω and θ would be positive in the Fixed phase (i.e., better cognitive mappers would stick with one strategy more often than non-cognitive mappers), but the correlation would become negative in the Random phase (i.e., cognitive mappers may be more flexible in how they approach spatial problems). In this way, the hybrid RL model allows us to test a very specific but important prediction about the cognitive basis of human navigation performance. Finally, to show that ω was indeed reflective of spatial navigation ability, we correlated ω with objectively measured wayfinding performance, with the hypothesis that these two factors were significantly correlated.

To foreshadow our results, we found that the model-free model outperformed the model-based model in the Fixed phase, but vice versa in the Random phase. The hybrid model, on the other hand, was the best model of human navigation in both phases. Participants relied on cognitive maps more and deviated from their default strategy more in the Random than in the Fixed phases. Supporting our theoretical framework, the correlations between model-based reliance and exploration–exploitation were different between the Fixed and Random phases. Lastly, wayfinding performance was correlated with model-based reliance.

Methods

Participants. One hundred and twenty-six participants from Georgia Institute of Technology and the Atlanta community participated in this experiment, either for course credits or monetary compensation. Participants spent between 80 to 140 min completing the experiment. Twelve participants felt motion sensitive and did not finish the experiment. As a result, one hundred and fourteen participants (forty-six females) were included in the data analysis. A sensitivity power analysis showed that the smallest effect size our study could detect was $r = 0.26$ given our final sample size (114), targeting statistical power (0.8) and alpha level (0.05), which was sensitive enough to detect small (0.2) to medium (0.5) effects according to Cohen's guidelines[51]. All participants (age 18–33) gave written consent and informed consent was obtained from all participants. The research was approved by the Institutional Review Board of Georgia Institute of Technology (IRB approval Code: H17456). All procedures were performed in accordance with the institutional guidelines.

Materials and procedure. *Navigation in virtual environment.* Participants completed a practice session in a 4 × 4 grid of rooms to familiarize with the control scheme and the objective of the navigation task. The 3D virtual environment was created in Sketchup (www.sketchup.com) and the navigation task was rendered and implemented in Unity 3D video game engine (https://unity.com/). Each room was a square of 10 × 10 virtual meters in size with a wall of 3 virtual meter. There was a penetrable door in each side of the room except for the rooms at the boundary. Movement in the virtual environment was enabled by keyboard, which provided self-paced and continuous translation and rotation. After the practice session, participants started the Fixed phase.

Fixed phase. To assess navigational learning and model it using RL algorithms, participants learned to navigate to hidden locations in a 6 × 6 grid of virtual rooms (Fig. 1). Each room had a unique reference object (toys, furniture, vehicles, etc.) served as local landmark which could only be seen within the room but not from other rooms. No distal or global landmark was available. Over the course of nine trials in the Fixed phase participants were instructed to find three specific goal objects repeatedly (apple, banana and watermelon; three trials per goal). These goal objects remained in the same rooms throughout the experiment but only the to-be-found goal object was invisible in a specific trial (e.g., all reference objects would be visible in their rooms, but if the goal object was "apple" in this trial, the banana would not appear even if participants traversed across the banana's

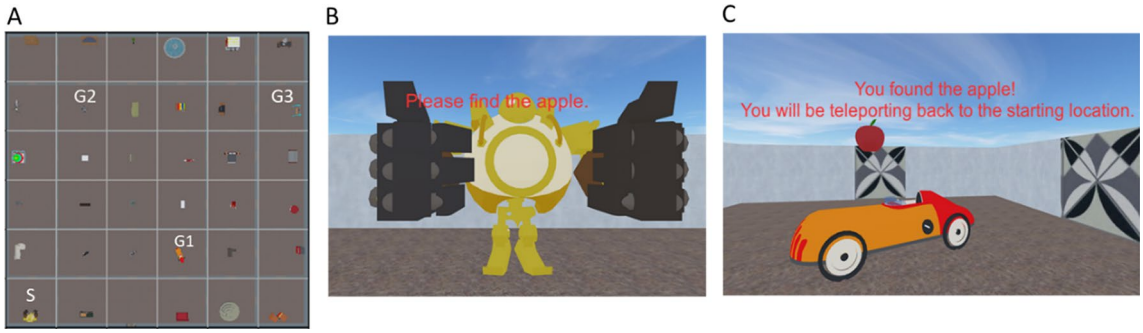

Figure 1. Experimental materials. (**A**) Layout of the environment. S indicates the fixed starting location in the Fixed phase, G1 ~ 3 indicate the goal object locations. (**B**, **C**) Actual view of landmark objects and rooms from the participants. Note that participants were brought back to the same starting location after finding a goal object during the Fixed phase.

room). This helped avoid blending learning of different goal-destination pairings in the same trial. Once participants had found the goal object, they were teleported to the starting location and were instructed to find the next goal object. To make this Fixed phase amenable to model-free learning, participants were always brought back to the same starting position with the same facing direction, and each goal object was to be found in the same order across participants (i.e., apple-banana-watermelon and then repeat for all participants). This is akin to learning the outbound paths from one's new home to the grocer, movie theater, etc. We limited the Fixed phase to nine trials to minimize the transfer of spatial learning, that is, many participants may start deriving shortcuts through model-based learning in the Fixed phase when they become extensively familiar with the environment[19,30,34], potentially suppressing our ability to delineate interesting individual differences.

Random phase. Participants then underwent a "Random phase" in the same virtual environment as in the Fixed phase. Importantly, our implementation of a small number of trials in the Fixed phase in our design not only ensured participants did not derive and habitualize "shortcuts" during the Fixed phase but also still had room to improve their precise configural knowledge of the environment and continue learning at this point. Therefore, the Random phase represented a critical period involving a probe of (1) spatial transfer and flexible perspective adoption from the Fixed phase and then (2) continued environmental learning under new procedures. The Random phase was almost identical to the Fixed phase except that participants' starting location and orientation were randomized in each trial (goal object locations were excluded in the possible starting locations). This is akin to finding the same grocer, movie theater, etc. from variable locations in the neighborhood. In addition, the order of goal objects was pseudorandomized such that each goal object was to be found once in every three trials but not in a predictable order (e.g., banana-apple-watermelon-apple-watermelon-banana…). There were seventy-two trials in the Random phase.

For our study, it was critical that the Fixed phase always preceded the Random for each participant. First, exposure to the Random phase prior to the Fixed phase may encourage participants to default to a model-based strategy and the performance in the Fixed phase could be at the floor level. Second, the Fixed phase—by repeating start-goal location pairings—enabled participants to develop one (or several) routes to a goal that would then be familiar in the Random phase and could be strategically exploited from a familiar landmark/room (enabling participants to exhibit shifts in strategy in Random where otherwise they would only have one [model-based] to go from).

Analyses pipeline outline. Described in detail below, our analysis pipeline was as follows: we first fit each participants' navigation behaviors with the three model-free models and selected the besting performing one (Fig. 3). We then created a hybrid model by combining the winning model-free model and a model-based model. Finally, we compared the performance of the model-free, model-based and hybrid models (Fig. 4) and chose the best performing model for subsequent individual differences analyses using its parameters (Table 1).

Reinforcement learning models. As mentioned in the Instruction, we used five different reinforcement learning models to fit navigation behaviors, separately for the Fixed and the Random phases and separately for each participant. We modelled the sequence of participants' choices (which rooms to enter) by comparing them step by step to those predicted by various models. As we had a 6 × 6 grid, the navigation task consisted of 36 states (rooms) and in each state, subjects could have up to four action choices (up, down, left or right). The navigation task consisted of three rewards (three goal objects), and the objective for all models was to learn the state-action value function Q(s,a) at each state-action pair (i.e., which direction to go when in a specific room to maximize reward) for each goal object (Fig. 2). We assumed no interference or generalization among the (implicit) rewards of the three goal objects, and thus each algorithm was subdivided into three independent task sets and value functions, one for each goal object.

Model-free reinforcement learning. To provide further insight of model-free behaviors in human spatial navigation and chose the best one for the hybrid model, we created three TD models: TD(0), TD(λ) and TD(1). We

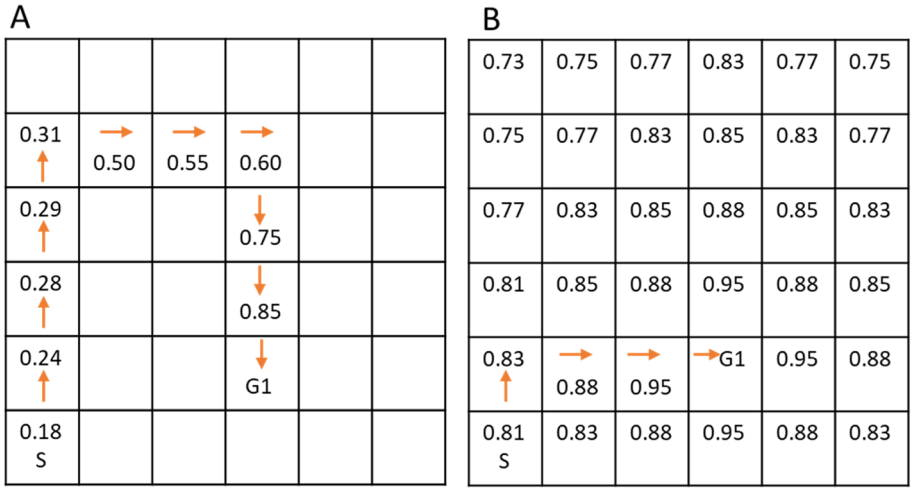

Figure 2. Model-free (**A**) and model-based (**B**) reinforcement learning models. The numbers in the figure are state values, showing how a navigator decides to move along the route in a given state/room. (**A**) Model-free valuation based on the TD algorithm. After finding an object this algorithm updates the values only along the traversed path. (**B**) Model-based valuations derived from dynamic programming. The model-based algorithm assumes a perfect cognitive map and the values in the entire environment are precomputed (See Model-based reinforcement learning). *S* starting location, *G1* Goal object # 1.

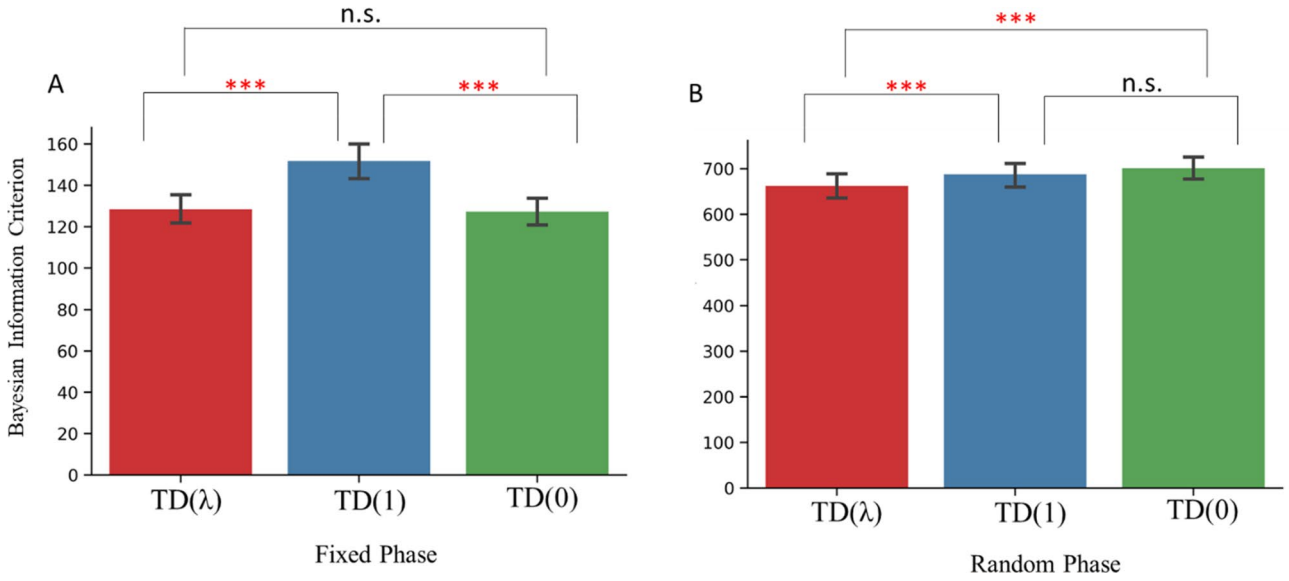

Figure 3. Model comparison in the Fixed (**A**) and the Random (**B**) phases. *BIC* Bayesian Information Criterion. *n.s.* not significant. ***p < 0.001.

first describe and provide the equations for TD(0), and then explain the differences between the three models. The equations for how the Q values were updated in the TD model were as follows[5]:

$$Q_{TD}(s_t, a_t) = Q_{TD}(s_t, a_t) + \alpha\delta \quad (1)$$

$$\text{where}\delta = r_{t+1} + Q_{TD}(s_{t+1}, a_{t+1}) - Q_{TD}(s_t, a_t) \quad (2)$$

Here, t denoted the current state and action, and $t+1$ denoted the future state and action chosen by the soft-max function (see below). Equation (1) showed that the Q value associated with the current state ($Q(s_{(t)}, a_{(t)})$) was updated by an error δ, adjusted by the learning rate α. Equation (2) showed that the error δ was determined by the reward associated with the future state ($r_{(t+1)}$) plus the difference between the Q values associated with the future and current states. The Q value of each state-action pair was initialized to be 0 in the beginning of the experiment, and the Q values were carried across trials and phases.

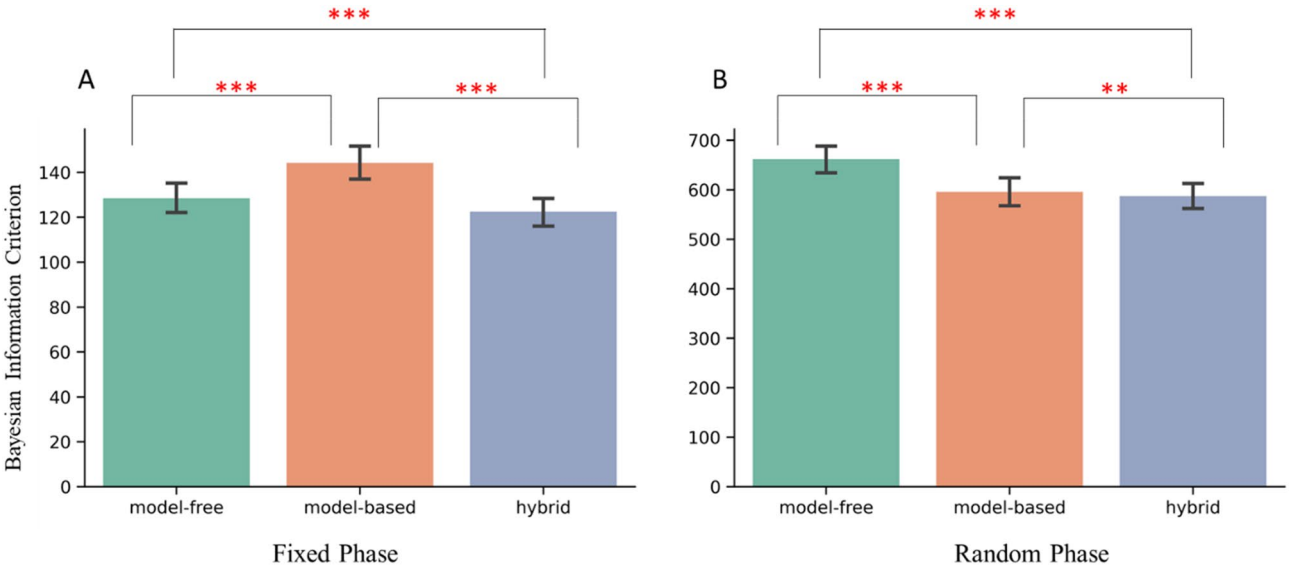

Figure 4. Model comparison in the Fixed (**A**) and the Random (**B**) phases. *BIC* Bayesian Information Criterion. **p < 0.01, ***p < 0.001.

	1	2	3	4
1. Fixed ω	–			
2. Random ω	0.53***	–		
3. Fixed θ	0.25**	0.31***	–	
4. Random θ	− 0.27**	− 0.35***	− 0.20*	–

Table 1. Correlation matrix between navigation strategy (ω) and exploration–exploitation (θ) in the Fixed and Random phases. *p < 0.05, **p < 0.01, ***p < 0.001.

To determine which action to take based on Q values associated with the future states and actions, we computed the probability of the action selection based on the softmax function:

$$p_{t+1} = \frac{exp^{\theta Q(s_{t+1},a)}}{\sum_{a' \in A} exp^{\theta Q(s_{t+1},a')}} \tag{3}$$

θ was the inverse temperature controlling the degree of randomness in participants' action selection, and a′ denoted the possible future actions from the current state. θ was constrained between 1 to 15 based, and the higher the θ, the more deterministic of the action selection and therefore more exploitative.

Compared to TD(0), TD(λ) added the eligibility trace *e* to the Q value updating[5], which was a temporary record of how frequently each state was visited. The eligibility trace for each state-action pair was set to be 0 in the beginning of each trial. The equations for how the Q values were updated in TD(λ) were as follows:

$$Q_{TD(\lambda)}(s_t, a_t) = Q_{TD(\lambda)}(s_t, a_t) + \alpha \delta e_t(s, a) \tag{4}$$

$$\text{where } e_t(s, a) = \lambda e_{t-1}(s, a) + \mathbf{I}(S_t = s, A_t = a) \tag{5}$$

I was the indicator function, which was equal to 1 when the condition inside it was true and 0 otherwise. λ was constrained between 0 and 1, so Eq. (5) indicated that the less frequently a state was visited, the smaller the updating of the Q value associated with that state. TD(1) was a special case of TD(λ), which forced every visited state to get the same amount of updating regardless of how often they were visited. When relating these RL algorithms to human memory systems, TD(0) assumed that memory updating, which was represented in the Q value updating, occurred only in the most recent visited location, whereas TD(λ) assumed that memory updating occurred in all previously visited locations continuously and such updating scaled with the frequency of visitation. TD(1) differed from TD(λ) that memory updating did not scale with the frequency of visitation.

Model-based reinforcement learning. For the model-based algorithm, we used dynamic programing[5] which learned the layout of the environment (i.e., cognitive map) by computing the Q values via traversing all possible rooms and directions to locate the goal (Fig. 2). We computed the Q values based on a 'sweeping' process termi-

nating at goal locations. We first initialized all $Q_{MB}(s)$ to 0 in the beginning of the Fixed phase. Then, for all states and adjacent state-action pairs we iteratively performed the following:

$$Q_{MB}(s) \leftarrow \sum_a \pi(a|s) + \sum_{s',r} p(s', r|s, a)\left[r + \gamma Q_{MB}(s')\right] \tag{6}$$

where $(a|s)$ was the probability to take action a from state s following the exploration vs. exploitation policy. $p\left(s', r|s, a\right)$ was the probability to end up in state s' and receive reward r given the current state and action. The algorithm had one fixed parameter γ set at 0.8. The final model-based values (Q_{MB}) were the values after the algorithm converged (i.e., the difference between each of the Q_{MB} in the current iteration and the previous iteration was smaller than 0.0001). Conceptually, the model-based values reflected the state-action values as if one had a perfect cognitive map, and therefore the Q values did not get updated and they were the same for all participants.

Hybrid model. We implemented a hybrid model as a weighted linear combination of the values from the best performing model-free algorithm across participants and the model-based algorithm:

$$Q_{hybrid} = (1 - \omega)Q_{MF} + \omega Q_{MB} \tag{7}$$

where ω represented the balance between the model-free and model-based behaviors. The higher the ω, the better the navigation behaviors could be characterized as model-based or cognitive map guided.

Model fitting and comparison. For each algorithm, we computed the negative log-likelihood (NLL) of the observed choices (a_t) by the summing over the log of Eq. (3), for the action chosen on each of the n trials, as follows:

$$NLL(\boldsymbol{X}) = -\sum_{t=1}^{n} logp(a_t|\boldsymbol{X}) \tag{8}$$

where vector **X** denotes the free parameters of the model, and the NLL was computed separately in Fixed and Random phases. The best fitting parameters were then computed as those that minimize the negative log likelihood:

$$\boldsymbol{X}_{\mathrm{MLE}} = arg\min_x NLL(\boldsymbol{X}) \tag{9}$$

Model fitting was performed using the the optimization function from SciPy[52]. Model comparison was performed by computing the Bayes Information Criterion (BIC) for each model for each participant, separately in the Fixed and Random phases.

$$BIC = klogn + 2\boldsymbol{X}_{\mathrm{MLE}} \tag{10}$$

where k is the number of free parameters in the model and n is the number of trials in the data. There were two free parameters (α and θ) in the TD(0) and TD(1) models, and three free parameters (α, θ and λ) in TD(λ). There was one free parameter θ in the model-based model, and four free parameters (α, θ, λ and ω) in the hybrid model.

Excessive distance. Excessive distance (ED) has been a widely used index to indicate wayfinding efficiency[19,22,49,53] which was defined as:

$$(\text{actual traversed distance} - \text{optimal distance})/\text{optimal distance}.$$

An ED of 0 indicated perfect wayfinding performance (actual traversed distance equals to optimal distance) and an index of 1 indicated the actual traversed distance was 100% longer than the optimal distance. In our study, because states and state transitions were compartmentalized by rooms, we used the number of rooms to represent distance.

Results

We used JASP (JASP Team, 2021) and Cocor[54] for statistical analyses, and Matplotlib[55] and Seaborn for data visualization.

TD(λ) outperformed other TD models in fitting spatial navigation behavior. We first compared the TD family algorithms in modeling navigation behavior in the Fixed and Random phases, separately (Fig. 3). In the Fixed phase, the one-way repeated ANOVA, with the three TD models as independent variable and BIC as dependent variable, was significant ($F(2,226) = 49.77$, $p < 0.001$, $\eta^2 = 0.30$). Paired t-test showed that TD(λ) outperformed TD(1) ($t(113) = -7.04$, $p < 0.001$, Cohen's $d = -0.66$), and was similar to the TD(0) model ($t(113) = 0.94$, $p = 0.35$, Cohen's $d = 0.09$). The TD(0) model outperformed the TD(1) model ($t(113) = -7.59$, $p < 0.001$, Cohen's $d = -0.711$; Fig. 3A). In the Random phase, the one-way repeated ANOVA was also significant ($F(2,226) = 10.48$, $p < 0.001$, $\eta^2 = 0.09$). Paired t-test showed that TD(λ) outperformed TD(1) ($t(113) = -3.62$, $p < 0.001$, Cohen's $d = -0.34$), and the TD(0) model ($t(113) = -4.52$, $p < 0.001$, Cohen's $d = -0.42$). There was no significant difference between the TD(0) and the TD(1) models ($t(113) = 1.36$, $p = 0.18$, Cohen's $d = 0.13$; Fig. 3B).

Overall, the TD(λ) was the best performing model among our selection of model-free models, and therefore we used it for the hybrid model.

The hybrid model outperformed the model-free and model-based models in both phases. We next compared which model, namely the model-free, model-based and the hybrid, provided the best fit in the Fixed and Random phases, respectively (Fig. 4). In the Fixed phase, the one-way repeated ANOVA was significant ($F(2,226) = 105.09$, $p < 0.001$, $\eta^2 = 0.48$). Paired t-test showed that the hybrid model outperformed the model-free ($t(113) = -4.07$, $p < 0.001$, Cohen's $d = -0.38$), and the model-based models ($t(113) = -17.93$, $p < 0.001$, Cohen's $d = -1.68$). Furthermore, the model-free model outperformed the model-based model ($t(113) = -8.34$, $p < 0.001$, Cohen's $d = -0.78$; Fig. 4A). In the Random phase, the one-way repeated ANOVA was also significant ($F(2,226) = 114.74$, $p < 0.001$, $\eta^2 = 0.50$). Paired t-test showed that the hybrid model outperformed the model-free ($t(113) = -11.62$, $p < 0.001$, Cohen's $d = -1.09$), and the model-based models ($t(113) = -3.07$, $p = 0.003$, Cohen's $d = -0.29$). Contrary to the results in the Fixed phase, the model-based model outperformed the model-free model in the Random phase ($t(113) = -10.59$, $p < 0.001$, Cohen's $d = -0.99$; Fig. 4B). Clearly, the hybrid model the best performing model in both navigational phases.

The correlation between navigation strategy (ω) and exploration–exploitation tendency (θ) was modulated by navigation requirement. Lastly, we asked whether navigation strategy (ω) and exploration–exploitation (θ) was modulated by navigation requirements. ω was significantly smaller in the Fixed than in the Random phases ($t(113) = -17.56$, $p < 0.001$, Cohen's $d = -1.64$), suggesting that in general, participants' navigation behaviors reflected more model-free in the Fixed than in the Random phases. On the other hand, θ was significantly larger in the Fixed than in the Random phases ($t(113) = -7.75$, $p < 0.001$, Cohen's $d = 0.73$), suggesting that in general, participants used the same navigation strategy more consistently in the Fixed than in the Random phases. We then compared the correlations ω and θ in the Fixed and Random phases (Table 1). In the Fixed phase, this correlation was significantly positive ($r(114) = 0.25$, $p = 0.007$), suggesting that in the Fixed phase, the cognitive mappers tended to exploit or used the same navigation strategy more consistently than the route followers. In the Random phase, on the other hand, this correlation became significantly negative ($r(114) = -0.35$, $p < 0.001$), suggesting that in the Random phase, the cognitive mappers tended to explore or vary their navigation strategy more than the route followers. Together, these results supported our theoretical framework of one important way that cognitive mappers differ from route followers: cognitive mappers are flexible and efficient not only by virtue of making use of cognitive map-based strategies, but by adaptively avoiding or embracing strategy change based on different navigational requirements.

To demonstrate that ω was also correlated with objectively observed performance, we correlated ω and excessive distance (ED). We found that ω was correlated with ED significantly in both Fixed and Random phases ($rs(114) < -0.51$, $ps < 0.001$), supporting the prediction that regardless of navigational requirements, more model-based behavior is indicative of being a better, more spatially-efficient navigator.

Discussion

The current study compared five RL models in characterizing human behaviors in navigation tasks with different requirements, and we found that a hybrid model, consisting of both model-free and model-based learning, provided the best fit in both navigation tasks, despite being penalized (in model comparison) for its greater complexity. Furthermore, through individual differences analyses, we found that the reliance on the model-based system (ω) and the variability of using the default strategy (θ) increased as the randomness of the wayfinding increased. Interestingly, the correlation between ω and θ was modulated by task requirements, such that individuals who relied more on model-based learning were more likely to stick with one navigation strategy when wayfinding was from the same starting location, but were more likely to vary their navigation strategy when wayfinding was from an unpredictable starting location.

We first compare three model-free models, namely the TD(0), TD(λ) and TD(1), to determine the role of memory updating in spatial navigation. As mentioned in the Introduction and Methods, TD(0) assumes memory updating only occurs in the most recent visited location. On the other hand, the eligibility trace in TD(λ) assumes that memory updating occurs in all previously visited locations and the amount of updating decreased over time if such locations were not visited again. TD(1) is the special case of TD(λ) that memory updating is the same in all previously visited locations regardless of their visitation frequency. Our results show that although the TD(λ) model is not better than the TD(0) model in the Fixed phase, it does outperform the other model-free models of our navigators' cognition in the Random phase. As evidenced by the superiority of the hybrid model in this phase over a purely model-based approach, and by virtue of TD(λ)'s properties, these findings suggest that to the extent that people exhibit TD-like profiles their spatial memory updating typically occurs in a more continuous manner across all previously visited locations and scales with visitation frequency in spatial navigation, especially when wayfinding is not completely deterministic (i.e., the Random phase). Our findings not only complement the literature on memory updating in spatial navigation[33,42–46], but also extend these findings via a computational approach.

As stated in the Introduction, the increasing familiarity of the environment and the demands of the Random phase would encourage participants to rely on map-like knowledge in a greater extent in the Random phase. Indeed, when compared the performance of model-free learning against model-based learning, we find that the model-free learning outperforms model-based in the Fixed phase, but is outperformed by model-based in the Random phase, which validates our modeling methods. The hybrid model, on the other hand, outperforms the model-free and model-based models in both learning phases, suggesting that the majority of the individuals did not entirely rely on either the model-free or model-based learning systems, in either scenario, but instead fell

somewhere in between. These findings are aligned with the well-established findings of the substantial individual differences in spatial navigation, such that—although some individuals have little or near perfect configural knowledge of their environment—most of fall somewhere in between on various objective measures[22–24,26,36,56]. To the best of our knowledge, this is the first study showing that a hybrid RL model significantly outperforms the model-free and model-based models in human spatial navigation explicitly.

After confirming that the results from the hybrid model were in line with the findings in the literature, we extracted the two key parameters to give a new understanding of individual differences in spatial navigation in such tasks. A large portion of the literature in human spatial navigation investigates what makes a good navigator[57], and a critical component is navigation strategy (route-following or cognitive mapping)[29–31,34]. As mentioned in the "Introduction", the most widely used method of measuring one's navigation strategy is the dual-solution task, which may simplify the complex navigation behaviors observed in humans. The parameter ω of the hybrid model indicates the proportion which people use a cognitive map in spatial navigation relative to following familiar routes, and it is significantly correlated with, but importantly is not identical to, navigation performance. The critical difference between ω and the solution index from the dual-solution task is that the solution index indicates a general navigation strategy as a trait over a number of trials, but ω can be used to indicate a general strategy (like what we did in the current study), but also be used in a trial-by-trial basis where each trial has its own ω. In other words, the RL models can provide a finer grained measure of one's navigation strategy compared to traditional methods.

Another important and unique contribution of the RL models is the parameter of the exploration–exploitation tendency (θ), which reflects how consistently one uses their default strategy and is very difficult to estimate with traditional methods. The combination of ω and θ provides a powerful and unique way to further understand what contribute to the substantial individual differences in spatial navigation. For example, in our study the change in correlation between ω and θ reveals that cognitive mappers use different strategies more often when the randomness in the task increases, which is a novel finding on what makes a good navigator[57]. This finding is also important because although model-based behavior itself captures navigational flexibility, we also see that people with good cognitive maps and an ability to engage in model-based behavior are also the same people who more often appropriately shift between frameworks For example, imagine rounding a corner and a familiar set of cues becomes visible—it may be that a familiar path forward from this point, aligning with model-free behavior, is in fact the most efficient option, and our computational approach reveals how a good cognitive mapper may make this switch. Taken together, these findings not only give us a new perspective for understanding the individual differences in human spatial navigation ability, through the lens of navigation strategies and the consistency of using such strategies, but also have important implications for how one might improve spatial navigation ability in humans as well as artificial agents.

Limitations

In the current study, the Fixed phase always preceded the Random phase. As elaborated in the "Methods", this was particularly important for the design and current research questions—however, an obvious step for subsequent studies would be to leverage a longer design that includes switches back to Fixed trials (either interleaved with Random or blocked) at different stages of learning, in order to understand how the experience of navigating our environment in more flexible ways (Random) may influence our internal model used when navigating more familiar relationships between locations.

Another important design consideration was that the environment layout was a regular grid-shape without any global landmarks. These attributes combine to make orientation in the global space more difficult and dependent on learning relationships between adjacent rooms. On the one hand, we view this as a strength of the design, especially in the context of studying state-to-state associations in a reinforcement learning model. And indeed, there are many real-world scenarios that share features with our virtual environment—navigating enclosed spaces, such as the interior of a hotel, mall, or hospital, and to a lesser extent subway maps. In addition, when nestled down among tall buildings in city streets our view of landmarks are more likely to be constrained to vista space and a local scale[58], somewhat akin to the present task's constraints (but not wholly). Nonetheless, another obvious step forward from this research is to investigate whether the shape of the environment and the presence of global landmarks affects the pattern of results reported in the current study. One might hypothesize that—by facilitating shortcutting—global landmarks would reduce the propensity of some people to exploit familiar route segments in lieu of exploration at the Random phase transition. On the other hand, pivoting from some of our other recent findings[22], it may also be the case that better navigators are those who flexibly take advantage of these additional cues more-so than poorer navigators, in which case an intriguing prediction from the present findings is that such cues may exacerbate differences between more and less-model-based individuals, and their variability in strategy, according to both task demands and cue availability. Future studies could test these complementary ideas.

Conclusions

In the current study we compare five reinforcement learning models in fitting human spatial navigation behaviors. We find that the hybrid model provides the best fit to the data regardless of task requirements, and it sheds important light on how task requirement modulates the navigation strategy (the balance between model-free and model-based), the consistency of using, and the interaction between these two factors. All in all, we show that reinforcement learning models provide a finer grained characterization of navigation strategy in a continuous manner based on individual's prior navigation history, which not only complements and extends the existing methods of studying individual differences in spatial navigation, but also suggests that the consistency of using a navigation strategy based on navigation requirements is an important factor of what makes a good navigator.

Data availability

The data that support the findings of this study and the analysis code are available from the corresponding authors upon request.

Received: 4 April 2022; Accepted: 8 August 2022
Published online: 17 August 2022

References

1. Collins, A. G. E. Reinforcement learning: Bringing together computation and cognition. *Curr. Opin. Behav. Sci.* **29**, 63–68 (2019).
2. Eckstein, M. K., Wilbrecht, L. & Collins, A. G. What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Curr. Opin. Behav. Sci.* **41**, 128–137 (2021).
3. Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annu. Rev. Psychol.* **68**, 101–128 (2017).
4. Lockwood, P. L. & Klein-Flügge, M. C. Computational modelling of social cognition and behaviour: A reinforcement learning primer. *Soc. Cogn. Affect. Neurosci.* https://doi.org/10.1093/scan/nsaa040 (2020).
5. Sutton, R. S. & Barto, A. G. *Reinforcement Learning, Second Edition: An Introduction* (MIT Press, 2018).
6. Thorndike, E. L. Animal intelligence: An experimental study of the associative processes in animals. *Psychol. Rev. Monogr. Suppl.* **2**, 1–109 (1898).
7. Eckstein, M. K. & Collins, A. G. E. Computational evidence for hierarchically structured reinforcement learning in humans. *PNAS* **117**, 29381–29389 (2020).
8. Collins, A. G. E. & Frank, M. J. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur. J. Neurosci.* **35**, 1024–1035 (2012).
9. Otto, A. R., Gershman, S. J., Markman, A. B. & Daw, N. D. The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol. Sci.* **24**, 751–761 (2013).
10. van de Vijver, I. & Ligneul, R. Relevance of working memory for reinforcement learning in older adults varies with timescale of learning. *Aging Neuropsychol. Cogn.* **27**(5), 654–676 (2019).
11. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
12. Simon, D. A. & Daw, N. D. Neural correlates of forward planning in a spatial decision task in humans. *J. Neurosci.* **31**, 5526–5539 (2011).
13. Jocham, G., Klein, T. A. & Ullsperger, M. Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. *J. Neurosci.* **31**, 1606–1613 (2011).
14. Vikbladh, O. M. *et al.* Hippocampal contributions to model-based planning and spatial memory. *Neuron* **102**, 683–693 (2019).
15. Schultz, W. Behavioral dopamine signals. *Trends Neurosci.* **30**, 203–210 (2007).
16. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
17. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
18. Anggraini, D., Glasauer, S. & Wunderlich, K. Neural signatures of reinforcement learning correlate with strategy adoption during spatial navigation. *Sci. Rep.* **8**, 10110 (2018).
19. He, Q., McNamara, T. P., Bodenheimer, B. & Klippel, A. Acquisition and transfer of spatial knowledge during wayfinding. *J. Exp. Psychol. Learn. Mem. Cogn.* **45**, 1364–1386 (2019).
20. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *PNAS* **110**, 20941–20946 (2013).
21. Radulescu, A., Daniel, R. & Niv, Y. The effects of aging on the interaction between reinforcement learning and attention. *Psychol. Aging* **31**, 747–757 (2016).
22. He, Q., McNamara, T. P. & Brown, T. I. Manipulating the visibility of barriers to improve spatial navigation efficiency and cognitive mapping. *Sci. Rep.* **9**, 1–12 (2019).
23. He, Q., Han, A. T., Churaman, T. A. & Brown, T. I. The role of working memory capacity in spatial learning depends on spatial information integration difficulty in the environment. *J. Exp. Psychol. Gen.* **150**, 666–685 (2021).
24. He, Q., Beveridge, E. H., Starnes, J., Goodroe, S. C. & Brown, T. I. Environmental overlap and individual encoding strategy modulate memory interference in spatial navigation. *Cognition* **207**, 104508 (2021).
25. Chrastil, E. R. & Warren, W. H. Active and passive spatial learning in human navigation: Acquisition of survey knowledge. *J. Exp. Psychol. Learn. Mem. Cogn.* **39**, 1520–1537 (2013).
26. Ishikawa, T. & Montello, D. R. Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. *Cogn. Psychol.* **52**, 93–129 (2006).
27. Weisberg, S. M., Schinazi, V. R., Newcombe, N. S., Shipley, T. F. & Epstein, R. A. Variations in cognitive maps: Understanding individual differences in navigation. *J. Exp. Psychol. Learn. Mem. Cogn.* **40**, 669–682 (2014).
28. Hegarty, M., Richardson, A. E., Montello, D. R., Lovelace, K. & Subbiah, I. Development of a self-report measure of environmental spatial ability. *Intelligence* **30**, 425–447 (2002).
29. Boone, A. P., Gong, X. & Hegarty, M. Sex differences in navigation strategy and efficiency. *Mem. Cogn.* https://doi.org/10.3758/s13421-018-0811-y (2018).
30. Marchette, S. A., Bakker, A. & Shelton, A. L. Cognitive mappers to creatures of habit: Differential engagement of place and response learning mechanisms predicts human navigational behavior. *J. Neurosci.* **31**, 15264–15268 (2011).
31. Boone, A. P., Maghen, B. & Hegarty, M. Instructions matter: Individual differences in navigation strategy and ability. *Mem. Cogn.* https://doi.org/10.3758/s13421-019-00941-5 (2019).
32. Kuliga, S. F. *et al.* Exploring individual differences and building complexity in wayfinding: The case of the Seattle central library. *Environ. Behav.* https://doi.org/10.1177/0013916519836149 (2019).
33. He, Q. & McNamara, T. P. Spatial updating strategy affects the reference frame in path integration. *Psychon. Bull. Rev.* **25**, 1073–1079 (2018).
34. Brown, T. I., Gagnon, S. A. & Wagner, A. D. Stress disrupts human hippocampal-prefrontal function during prospective spatial navigation and hinders flexible behavior. *Curr. Biol.* **30**(10), 1821–1833 (2020).
35. Brown, T. I., Whiteman, A. S., Aselcioglu, I. & Stern, C. E. Structural differences in hippocampal and prefrontal gray matter volume support flexible context-dependent navigation ability. *J. Neurosci.* **34**, 2314–2320 (2014).
36. He, Q. & Brown, T. I. Heterogeneous correlations between hippocampus volume and cognitive map accuracy among healthy young adults. *Cortex* **124**, 167–175 (2020).
37. Chrastil, E. R., Sherrill, K. R., Aselcioglu, I., Hasselmo, M. E. & Stern, C. E. Individual differences in human path integration abilities correlate with gray matter volume in retrosplenial cortex, hippocampus, and medial prefrontal cortex. *ENeuro* https://doi.org/10.1523/ENEURO.0346-16.2017 (2017).

38. Sherrill, K. R. *et al.* Functional connections between optic flow areas and navigationally responsive brain regions during goal-directed navigation. *Neuroimage* **118**, 386–396 (2015).
39. Bohbot, V. D., Lerch, J., Thorndycraft, B., Iaria, G. & Zijdenbos, A. P. Gray matter differences correlate with spontaneous strategies in a human virtual navigation task. *J. Neurosci.* **27**, 10078–10083 (2007).
40. Blacker, K. J., Weisberg, S. M., Newcombe, N. S. & Courtney, S. M. Keeping track of where we are: Spatial working memory in navigation. *Vis. Cogn.* **25**(7–8), 691–702 (2017).
41. Nazareth, A., Huang, X., Voyer, D. & Newcombe, N. A meta-analysis of sex differences in human navigation skills. *Psychon. Bull. Rev.* https://doi.org/10.3758/s13423-019-01633-6 (2019).
42. He, Q., McNamara, T. P. & Kelly, J. W. Reference frames in spatial updating when body-based cues are absent. *Mem. Cogn.* **46**, 32–42 (2018).
43. He, Q. & McNamara, T. P. Virtual orientation overrides physical orientation to define a reference frame in spatial updating. *Front. Hum. Neurosci.* **12**, 269 (2018).
44. Klatzky, R. L., Loomis, J. M., Beall, A. C., Chance, S. S. & Golledge, R. G. Spatial updating of self-position and orientation during real, imagined, and virtual locomotion. *Psychol. Sci.* **9**, 293–298 (1998).
45. Wang, R. F., Brockmole, J. R. & Abdul-Salaam, R. A. Spatial updating across environments. *J. Vis.* **2**, 420–420 (2002).
46. Wang, R. F. & Brockmole, J. R. Simultaneous spatial updating in nested environments. *Psychon. Bull. Rev.* **10**, 981–986 (2003).
47. Siegel, A. W. & White, S. H. The development of spatial representations of large-scale environments. *Adv. Child Dev. Behav.* **10**, 9–55 (1975).
48. Tolman, E. C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).
49. Newman, E. L. *et al.* Learning your way around town: How virtual taxicab drivers learn to use both layout and landmark information. *Cognition* **104**, 231–253 (2007).
50. Feng, S. F., Wang, S., Zarnescu, S. & Wilson, R. C. The dynamics of explore–exploit decisions reveal a signal-to-noise mechanism for random exploration. *Sci. Rep.* **11**, 3077 (2021).
51. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Routledge, 1988). https://doi.org/10.4324/9780203771587.
52. Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
53. He, Q., Han, A. T., Churaman, T. A. & Brown, T. I. The role of working memory capacity in spatial learning depends on spatial information integration difficulty in the environment. *J. Exp. Psychol. Gen.* https://doi.org/10.1037/xge0000972 (2020).
54. Diedenhofen, B. & Musch, J. cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE* **10**, e0121945 (2015).
55. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
56. Weisberg, S. M. & Newcombe, N. S. Cognitive maps: Some people make them, some people struggle. *Curr. Dir. Psychol. Sci.* https://doi.org/10.1177/0963721417744521 (2018).
57. Wolbers, T. & Hegarty, M. What determines our navigational abilities?. *Trends Cogn. Sci.* **14**, 138–146 (2010).
58. Wolbers, T. & Wiener, J. M. Challenges for identifying the neural mechanisms that support spatial navigation: The impact of spatial scale. *Front. Hum. Neurosci.* **8**, 571 (2014).

Author contributions

Q.H., J.L., L.E. and T.B. designed the experiment. J.L., L.E. and E.B. collected the data. Q.H. performed the data analysis. Q.H. wrote the main manuscript text and prepare the figures. All authors revised and reviewed the manuscript.

Funding

Funding was provided by Warren Alpert Foundation to Q.H. and National Institutes of Health (Grant No. 1-R21AG063131) to T.B.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.H. or T.I.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.