Predictive Maps in Rats and Humans for Spatial Navigation

William de Cothi[1 *], Nils Nyberg[2], Eva-Maria Griesbauer[2], Carole Ghanamé[2], Fiona Zisch[3], Julie M. Lefort[1], Lydia Fletcher[2], Coco Newton[4], Sophie Renaudineau[2], Daniel Bendor[2], Roddy Grieves[5], Éléonore Duvelle[5], Caswell Barry[1] & Hugo J. Spiers[2*]

1. Department of Cell and Developmental Biology, University College London, UK
2. Institute of Behavioural Neuroscience, Department of Experimental Psychology, Division of Psychology and Language Sciences, University College London, UK
3. The Bartlett School of Architecture, University College London, UK
4. Department of Clinical Neurosciences, University of Cambridge, UK
5. Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA

* Corresponding authors emails: w.decothi@ucl.ac.uk, h.spiers@ucl.ac.uk

Abstract

Much of our understanding of navigation comes from the study of individual species, often with specific tasks tailored to that species. Here, we provide a novel experimental and analytic framework, integrating across humans, rats and simulated reinforcement learning (RL) agents to interrogate the dynamics of behaviour during spatial navigation. We developed a novel open-field navigation task ('Tartarus Maze') requiring dynamic adaptation (shortcuts and detours) to changing obstructions in the path to a hidden goal. Humans and rats were remarkably similar in their trajectories. Both species showed the greatest similarity to RL agents utilising a 'successor representation', which creates a predictive map. Humans also displayed trajectory features similar to model-based RL agents, which implemented a tree-search planning procedure. Our results help refine models seeking to explain mammalian navigation in dynamic environments, and highlight the utility of modelling the behaviour of different species to uncover shared mechanisms that support behaviour.

Key words: Artificial Intelligence, Reinforcement Learning, Spatial Memory, Decision-making, Planning, Hippocampus, Behavioural Flexibility, Detour, Shortcut

Introduction

Adapting to change is fundamental for survival. Dealing with obstructions in terrain that require a detour are a common challenge for many animals, including humans navigating cities. The ability to infer an efficient detour or exploit a novel shortcut when it arises are considered hallmarks of the ability of animals to form a cognitive map of the environment (Epstein et al., 2017; Tolman, 1948; Tolman & Honzik, 1930). Both rodents and humans can show an impressive capacity to identify shortcuts and take optimal detours (Alvernhe et al., 2008; Brown et al., 2020; Grieves & Dudchenko, 2013; Howard et al., 2014; A. H. Javadi et al., 2017; A.-H. Javadi, Patai, Marin-Garcia, Margois, et al., 2019; Patai et al., 2019; Tolman et al., 1946; Tolman, 1948; Xu et al., 2010). However, not all studies show successful adaptive behaviour (Spiers & Gilbert, 2015). Rats often require multiple exposures to a set of paths before they are able to shift towards an optimal detour (Grieves & Dudchenko, 2013), and may fail to select an optimal shortcut from a set of novel paths (Gentry et al., 1947). Humans too have been shown to be poor at judging the directions between locations in walled mazes, hindering the capacity to identify shortcuts (Foo et al., 2005).

Much research into navigation implicitly assumes that rodents and humans navigate in a fundamentally similar way (Ekstrom & Ranganath, 2018; Epstein et al., 2017) and this has been used to support the integration of insights across both species (Gahnstrom & Spiers, 2020; O'Keefe & Nadel, 1978; Spiers & Barry, 2015; Tolman, 1948). In mammals, the hippocampus has been proposed to form a cognitive map of space enabling flexible navigation (O'Keefe & Nadel, 1978), evidenced by spatial-tuned cells (such as 'place cells') in the hippocampal formation of rodents and humans (Ekstrom et al., 2003; Grieves & Jeffery, 2017). However, despite the wide array of human and rodent research, few experiments have sought to compare rats and humans on a directly homologous task. Understanding the similarities and differences of these two species on the same task would be useful for allowing the better integration of findings from different methods, such as combining neuroimaging in humans with neural recordings and disruption methods in rodents (Barron et al., 2020, 2021). Moreover, such integration has potential benefit for the translation of assessments in rodents to assessments in clinical trials for humans, where tests of spatial navigation have been a major tool in the assessment of drug compounds (Laczó et al., 2012).

When considering how humans and rats might differ for navigation, differences in sensory perception may be important. Whilst humans have binocular vision, they also have diminished olfactory function and lack the tactility of whiskers. Rats have a larger visual field of view and can move their eyes independently (Wallace et al., 2013). In terms of neuroanatomy, the prefrontal cortical regions associated with spatial planning differ greatly between rats and primates (Patai & Spiers, 2021; Uylings et al., 2003); while the hippocampus and surrounding structures associated with spatial representations are relatively similar (Andersen et al., 2006). Given these similarities and differences in anatomy, we might expect them to transcend into the navigation strategies employed by rats and humans.

One approach for identifying potential cross-species mechanisms underlying goal-directed behaviour is through comparison with reinforcement learning models (Banino et al., 2018; Botvinick et al., 2020; Dayan & Daw, 2008; Momennejad, 2020; Simon & Daw, 2011). Reinforcement learning is an area of machine learning that addresses the theoretical problem of how an artificial agent should act in an environment in order to maximise rewards. Specifically, it seeks to maximise the expected accumulated reward R yielded from its future states s_t, which is known as the value function V:

$$V(s) \;=\; \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \,\middle|\, s_0 = s\right] \tag{1}$$

The parameter γ is a discount factor that determines the timescale of how motivating future rewards are, such that for $\gamma < 1$ the agent exponentially discounts rewards in the more distant future (Sutton & Barto, 2018).

Solutions to reinforcement learning problems have traditionally been divided into two categories: model-based methods that afford the agent a model of the environment, used to decide actions via a planning procedure (Sutton & Barto, 2018), and model-free methods that learn from experience which actions lead to the most rewarding future (Sutton, 1988; Watkins & Dayan, 1992). Providing that the model implemented in a model-based algorithm contains an accurate depiction of the environment, model-based methods are typically able to respond quickly and optimally to environmental perturbations. However, the planning procedure - for example a tree search (Silver et al., 2016) - required to successfully harness the model brings with it computational complexity and overhead, particularly in large state spaces with deep transition structures such as navigating a city.

In contrast to model-based methods, model-free methods are generally more simple and computationally inexpensive through a reliance on temporal-difference (TD) learning rules (Sutton, 1988), however, this comes with a reduced flexibility to environmental changes. As such, model-free mechanisms are often associated with the formation of habits (Dolan & Dayan, 2013; Lee & Keramati, 2017). To achieve this simplicity, model-free methods usually operate using the state-action value function Q instead of the state value function V. This makes it possible to disassociate the value of different actions a in a particular state s, without the need to know how the states are interconnected.

$$Q(s,a) \;=\; \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \,\middle|\, s_0 = s,\; a_0 = a\right] \tag{2}$$

Whilst model-free and model-based methods appear to function at opposite ends of an algorithmic spectrum, intermediary methods do exist. One such algorithm that has increased in application recently is the successor representation (SR) (Dayan, 1993). The SR somewhat combines parts of model-free and model-based learning (Gershman, 2018; Russek et al., 2017) by using experience to learn a predictive map M between the states in

an environment. For a one-step state transition matrix T, the predictive map can be learned via TD learning and is equivalent to the discounted sum of future state transitions:

$$M = I + \gamma T + \gamma^2 T^2 + \ldots = \sum_{t=0}^{\infty} \gamma^t T^t \tag{3}$$

This discounting of transitions means M can be readily combined with a separately learned reward R associated with each state s in order to explicitly compute value (equation 1), thus negating the need for a complicated planning procedure in order to use the model to guide action selection.

$$V(s) = \sum_{s'} M(s, s') R(s') \tag{4}$$

The SR has been able to provide a good account of behaviour and hippocampal representations in humans (Bellmund et al., 2019; Brunec & Momennejad, 2019; Garvert et al., 2017; Momennejad et al., 2017) and rodents (de Cothi & Barry, 2020; Stachenfeld et al., 2017). The tasks often used to draw these comparisons with model-free, model-based and SR behaviours typically focus on small state-spaces, with 2-step transition structures – as such the extent of planning often requires choosing a single action. Furthermore, due to the conceptual nature of the underlying task space, translational research usually requires differing sensory implementations for humans (Daw et al., 2011) and rodents (Miller et al., 2017).

Here, we created a configurable open-field maze with a shifting layout (Tartarus Maze) to test the navigation of rats in physical space, humans in immersive VR and RL agents in a simulation. Using a range of analytic methods we investigated how rat and human spatial behaviours compare to model-free, model-based and SR reinforcement learners. We found strong similarity in the trajectories of rats and humans. Both rats and humans showed the greatest likelihood similarity to SR based RL agents, with humans also displaying trajectory features similar to model-based RL agents implementing a tree-search planning procedure.

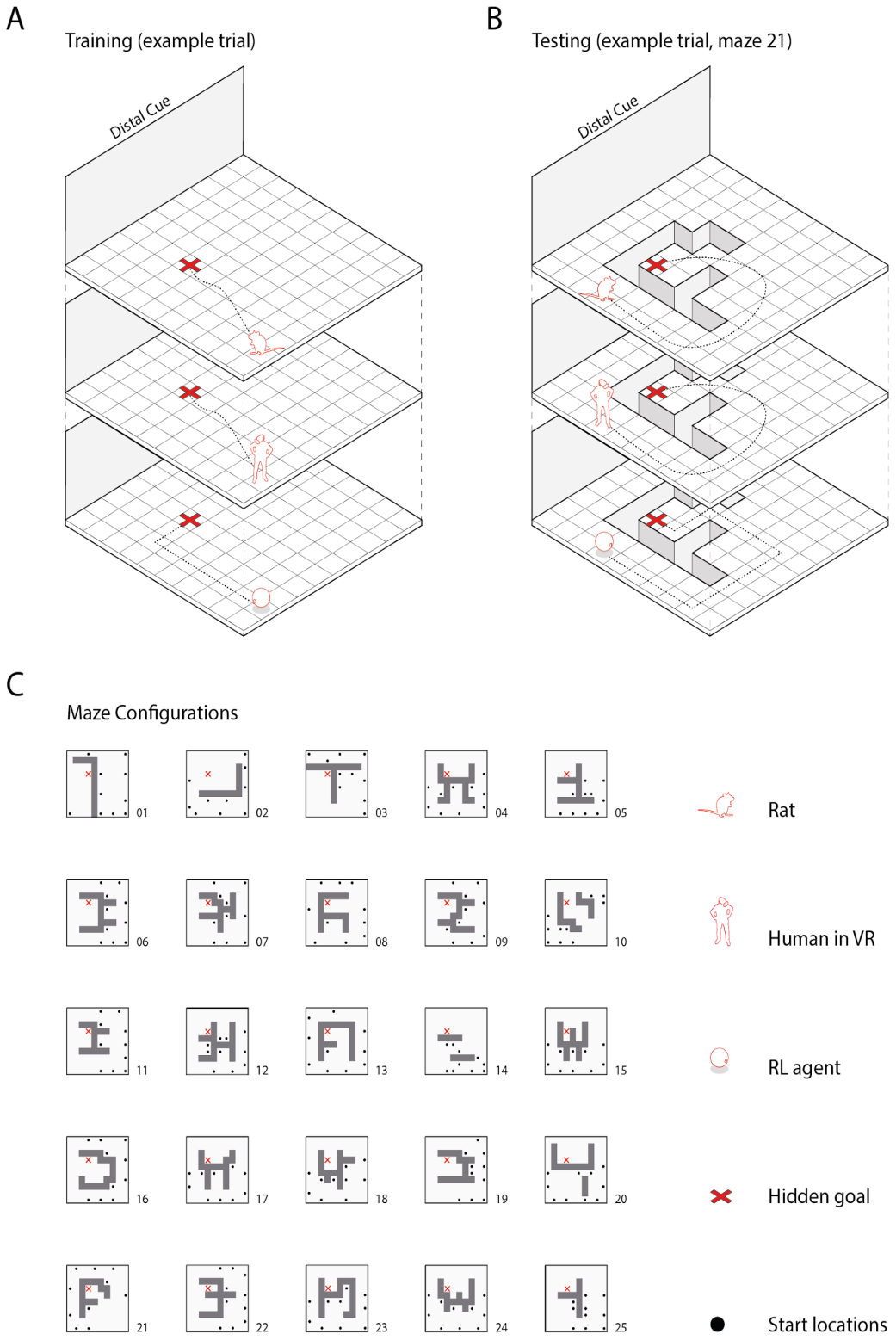


Figure 1: The Tartarus Maze. A) Schematic of the maze composed of 10x10 units for humans, rats and RL agents. For rats each unit (20x20cm) had a port to potentially deliver chocolate milk reward after the rat waited 5 sec at the goal (See Video S1). For humans each unit could be associated with a hidden gold star linked to financial reward, which appeared after waiting 5 sec at the goal location. Gaps between units were not visible to humans to avoid counting distance to the hidden goal (see Video S2). B) After training, flexible navigation was tested by removing units from the

maze to create gaps between traversable surfaces. Example configuration for trial 21 is shown. Each configuration was tested for 10 trials until 25 configurations had been tested. C) Sequence of 25 maze configurations used.

Results

Navigation was tested in a large square environment with a fixed hidden goal location and a prominent directional black wall cue in one direction (Fig. 1, Video S1 and S2). The maze was divided in a 10x10 grid of moveable sections that could either be removed, leaving impassable gaps to force detour taking, or added, creating shortcuts. During training, all 10x10 maze modules were present and the rats and humans were trained to reach the goal within a 45s time limit, (Fig 1A), while RL agents were initialised with the optimal value function. During the testing phase of the experiment, maze modules were removed to block the direct route to the goal (Fig 1B). Humans (n=18), rats (n=9) and agents were tested on the same sequence of 25 maze configurations each with 10 trials in which a set of defined starting locations were selected to optimally probe navigation (Fig 1C).

Behavioural performance is relatively similar between rats and humans

We first asked how well humans and rodents were able to complete the task during the test sessions. As expected, repeated exposure over trials to a new maze configuration corresponded to a general increase in the ability of both the humans and rats to navigate to the goal within the 45s time limit (Fig 2C; first 5 trials vs last 5 trials: humans $t(17)=6.3$, $p<.001$; rats $t(8)=4.0$, $p= .004$). Humans were also generally better than the rats at finding the goal across the 25 maze configurations (Fig 2D; humans vs rats: $t(25)=3.0$, $p=.006$). There were 3 mazes in which rats outperformed humans (mazes 2, 10 and 19). We saw a strong correlation between the places visited by the rats and human participants (occupancy correlation, humans vs rats: $\rho=.67$), in particular towards the later trials when both were better at navigating to the goal (Fig 2E; occupancy correlations for first 5 trials vs last 5 trials: $t(8)=3.2$, $p=.013$). The routes used were also more efficient (Fig 2E; deviation from optimal path, first 5 trials vs last 5 trials: human $t(17)=-5.0$, $p<.001$; rats $t(8)=-4.0$, $p=.004$) with humans generally choosing more optimal routes than the rats (deviation from optimal path humans vs rats: $t(25)=-8.2$, $p<.001$). Inspection of trajectories showed that in some cases near optimal paths could be observed in both species even on the first trial of a maze (see Video S1&2).

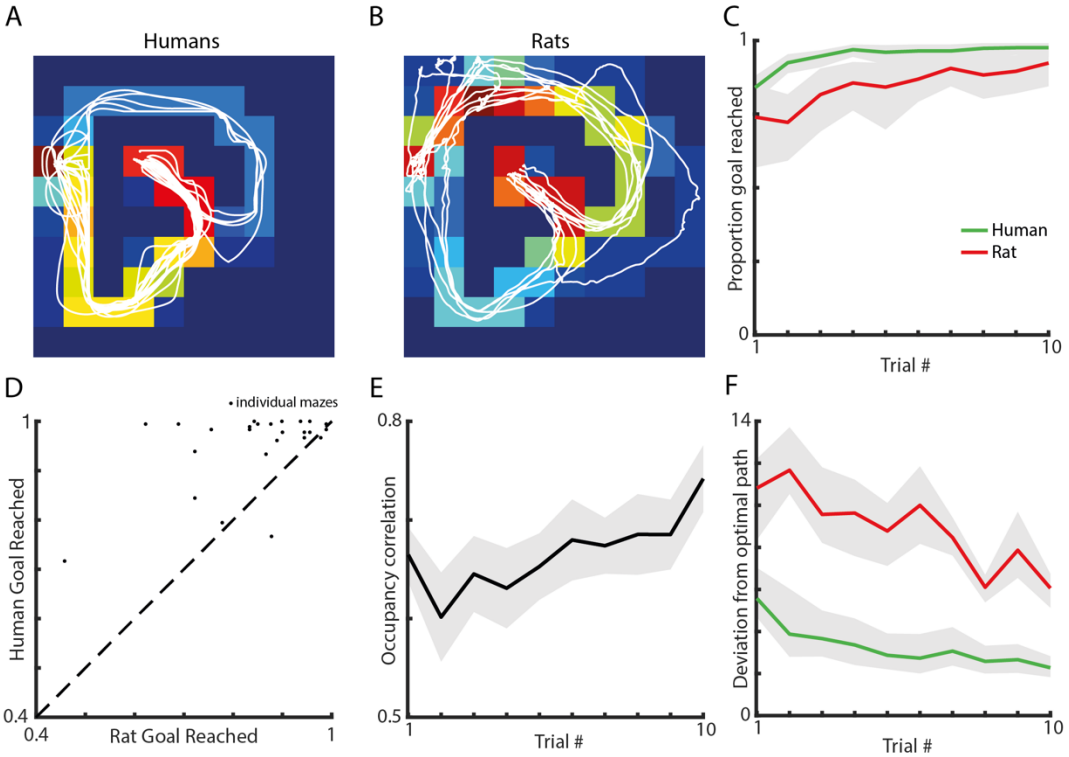


Figure 2: Humans and rats were able to successfully navigate to the hidden goal and generally did so using similar routes. Examples of the human (A) and rat (B) trajectories overlaying occupancy maps for a given trial. (C) Proportion of trials where the goal was reached averaged over all configurations as a function of trial number, for rats (red) or humans (green). Grey areas indicate 95% confidence intervals. D) Proportion of goals reached by humans and rats during the time limit. Configuration 2 is the only configuration where rats out performed humans. Configuration 1 was the configuration both species performed most inaccurately on. (E) Correlation between the human and rat occupancy maps. Note the increase across exposure to a maze configuration implying they take increasingly similar routes. (F) Average deviation from optimal path (measured in extra maze module visitations) for rats and humans as a function of trial number; humans navigate to the goal with more efficient routes.

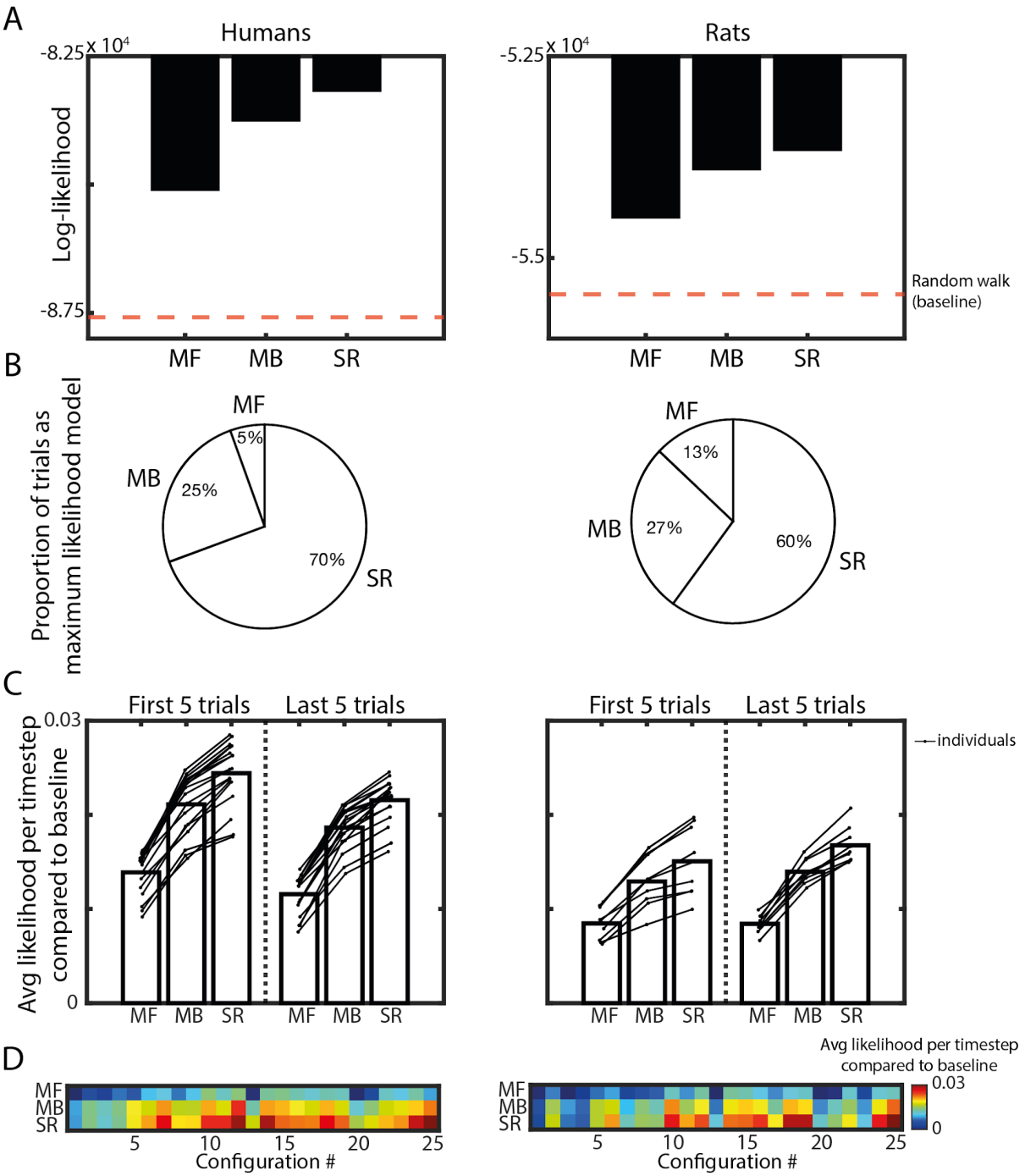
Likelihood analysis of actions reveals rats and humans are both most similar to an SR agent

We next investigated how the human and rat trajectories compared to the RL agents' trajectories. To compute the likelihood of the human and rat behaviour matching a specific model, we used their trajectories as input to the RL agents and used the agents' internal value estimates to compute a softmax probability distribution over the available actions at each timestep. Using these probabilities to compute the likelihood of the biological data underlying each agent, we calculated the maximum likelihood parameter estimates for each model's learning rate and discount factor across individuals, as well as pooled across the subject sample (Table 1).

	α		γ	
Model	Human	Rat	Human	Rat
MF	0.45	0.22	0.75	0.65
MB	1	1	0.81	0.79
SR	0.93	0.81	0.79	0.79

Table 1: Maximum likelihood parameters for learning rates α and discount factors γ for the model-free (MF), model-based (MB) and successor representation (SR) agents, pooled across individuals.

Comparing the model-free, model-based and SR algorithms, the value representation of the SR agent consistently provided the most likely fit to the biological behaviour (Fig 3A; Likelihood-Ratio (LR) test: SR vs MF for human data ln(LR) = 1911.1; SR vs MB for human data ln(LR) = 538.2; SR vs MF for rat data ln(LR) = 842.0; SR vs MB for rat data ln(LR) = 225.2), with the model-free agent consistently providing the worst fit (MF vs MB for human data ln(LR) = -1372.9; MF vs MB for rat data ln(LR) = -616.9). Consequently, the SR agent was the maximum likelihood model for 70% of the human trials and 60% of the rat trials (Fig 3B). Normalising these likelihoods by trial length and using a uniform random walk as a baseline, we observed this trend was robust throughout the time spent on a maze configuration (Fig 3C) and across individuals (SR vs MF for human data: $t(17) = 29.2$ $p<.001$; SR vs MB for human data: $t(17) = 11.9$, $p<.001$; SR vs MF for rat data: $t(8) = 13.0$, $p<.001$; SR vs MB for rat data: $t(8) = 9.6$, $p<.001$). We also observed that the agent likelihoods for humans and rats varied from maze-to-maze (Fig 3D), with a strong correlation between the fits to the biological data ($r=.57$, $p<.001$).



<u>Figure 3</u>: Maximum likelihood analyses of the human and rat trajectories. Likelihood analysis reveals that humans (left) and rats (right) are more similar to a successor representation agent (SR) than model-based (MB) or model-free (MF). (A) The value estimates generated by the SR agent provide a more likely explanation of the biological behaviour than either the MF or MB agents. (B) The SR agent was the maximum likelihood model to explain the biological behaviour for the majority of trials. (C) This trend is true across all individuals in both species (humans n=18; rats n=9) and robust throughout exposure to a maze. (D) Likelihood estimates vary from maze-to-maze, with a strong correlation between the model likelihoods for the human and rat behaviour across mazes ($r = .57$, $p<.001$).

Simulating agents using parameters derived from the human and rat data reveals closest match to SR agent

To investigate whether these differences in agent likelihoods transfer into differences in the resulting behavioural performance, we simulated the agent's trajectories using the pooled maximum likelihood parameters (Table 1) and an epsilon-greedy policy with epsilon=0.1. This policy ensures that agents exploit (i.e. choose the action with maximum expected value) 90% of the time and explore (i.e. choose a random action) on the remaining 10%. To accurately depict the distribution of trajectories generated by an agent under such a policy, we simulated each RL algorithm 1000 times, with the maximum number of state transitions set to the mean from the humans and rats on unsuccessful trials (humans: T = 42, rats: T = 29).

The biological behaviour consistently outperformed all RL agents. Beyond that, the model-based algorithm was the most successful agent (Fig 4A-B; Mann-Whitney test, human parameters: MB vs SR, z=29.6, p<.001; MB vs MF, z=38.7, p<.001; rat parameters: MB vs SR z=14.5 p<.001, MB vs MF z=38.7 p<.001), with the model-free agent consistently the worst (human parameters: MF vs SR, z=-38.7, p<.001; rat parameters: MF vs SR, z=-38.7 p<.001). As with the humans and rats, the model-based and SR agents progressively improved throughout the trials on a given maze configuration (Fig 4A-B; first 5 vs last 5 trials, human parameters: MB, z=-18.1, p<.001; SR, z=-30.1, p<.001; rat parameters: MB z=-14.2, p<.001; SR: z=-11.0, p<.001). Meanwhile, the model-free agents became progressively worse at reaching the goal (first 5 vs last 5 trials: human parameters, z=38.6, p<0.001; rat parameters, z=38.8, p<.001) – indicative of the increased tortuosity of the optimal trajectory. Goal-reaching performance for the RL algorithms varied from maze-to-maze (Fig C-D), with mazes that have a contradictory optimal policy to the preceding maze seeming particularly difficult (i.e. mazes 4, 8, 13, 21 – see Fig 1C for mazes). Conversely, mazes that possess a high degree of overlap in optimal policy with the previous maze (i.e. mazes 2, 7, 25) were consistent with higher levels of agent goal-reaching - presumably due to the improved accuracy of the initial value representations. Ranking maze difficulty by order of goal-reaching performance revealed a significant positive correlation between the human and rat difficulty rankings with the SR agent (Fig 4E-F; human-SR rank correlation: ρ=.52, p<.001; rat-SR rank correlation: ρ=.40, p=.049). No significant correlations were detected between the human and rat difficulty rankings with either of the other agents (human-MF rank correlation: ρ=.15, p=.48; human-MB rank correlation: ρ=-0.18, p=.39, rat-MF rank correlation: ρ=.37, p=.066; rat-MB rank correlation: ρ=.085, p=.69).

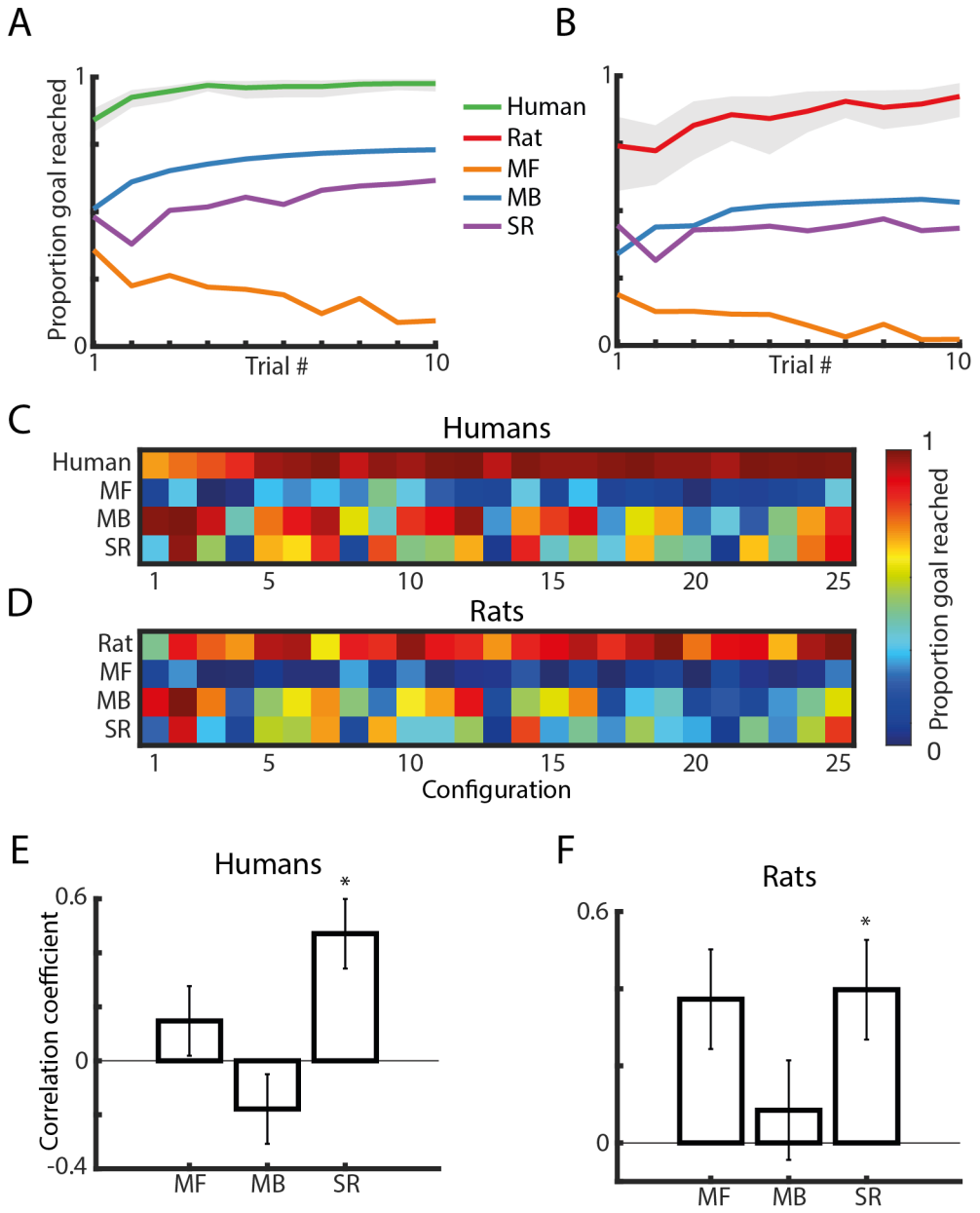
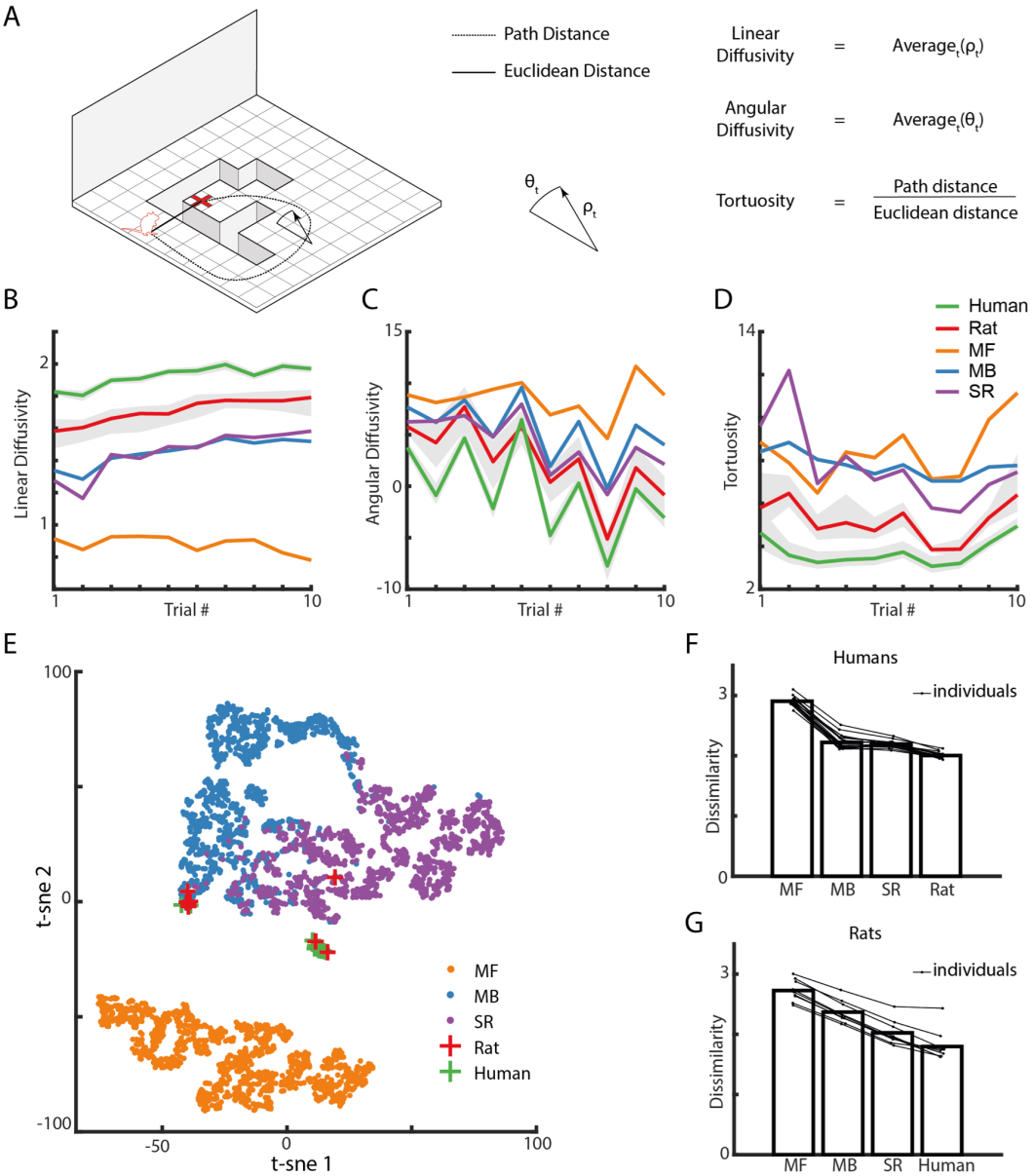


Figure 4: Human and rat performance compared with maximum likelihood reinforcement learning agents. Goal reaching for agents (n=1000) using the maximum likelihood parameters fit to the human (A) and rat (B) trajectories. Grey areas indicate 95% confidence intervals. (C-D) Goal reaching varied across maze configurations. Using this to rank maze configurations difficulty reveals a significant positive correlation between the human (E) and rat (F) difficulty rankings and the successor representation agent's (humans, ρ=.52, *p*<.001; rats, ρ=0.40, *p*=.049), but not with the model-free (humans,ρ=.15, *p*=.48; rats,ρ=.37, *p*=.066) or model-based (humans, ρ=-0.18, *p*=.39; rats, ρ=0.085, *p*=0.69) agent. This is in spite of the model-based agent generally being better at navigating to the goal than the other agents. Error bars indicate standard error from the mean.

In order to compare the similarity of the routes taken by the rats, humans and RL agents, we next quantified each trajectory using three measures: the linear diffusivity, the angular diffusivity and the tortuosity. (Fig 4A). Linear and angular diffusivity describe the

directness and direction of the route, whilst tortuosity describes the general trajectory complexity. In all measures, rats and humans values are closer to each other than to any of the reinforcement learners (Fig 5B-D).

Visualising the quantified trajectories for a specific trial revealed distinct clustering of behaviour (Fig 5E). We then used the Mahalanobis distance to calculate the dissimilarity between the biological and agent trajectories. The Mahalanobis distance was used as it accounts for the covariance between the trajectory measures when calculating dissimilarity. We found that human and rat trajectories were significantly more similar to each other than to any of the reinforcement learners (Fig 5FG; humans to rat cluster vs SR cluster: $t(17)=-26.3$, $p<.001$; rats to human cluster vs SR cluster: $t(8)=-6.0$, $p<.001$). Of the RL agents, the SR was the most similar to the rat (SR vs MF: $t(8)=-27.5$, $p<.001$; SR vs MB: $t(8)=-34.6$, $p<.001$) and human trajectories (SR vs MF: $t(17)=-97.7$, $p<.001$; SR vs MB: $t(17)=-2.4$, $p=.03$), with the model-free agent consistently being least similar (humans to MF cluster vs MB cluster: $t(17)=37.6$, $p<.001$; rats to MF cluster vs MB cluster: $t(8)=20.7$, $p<.001$).

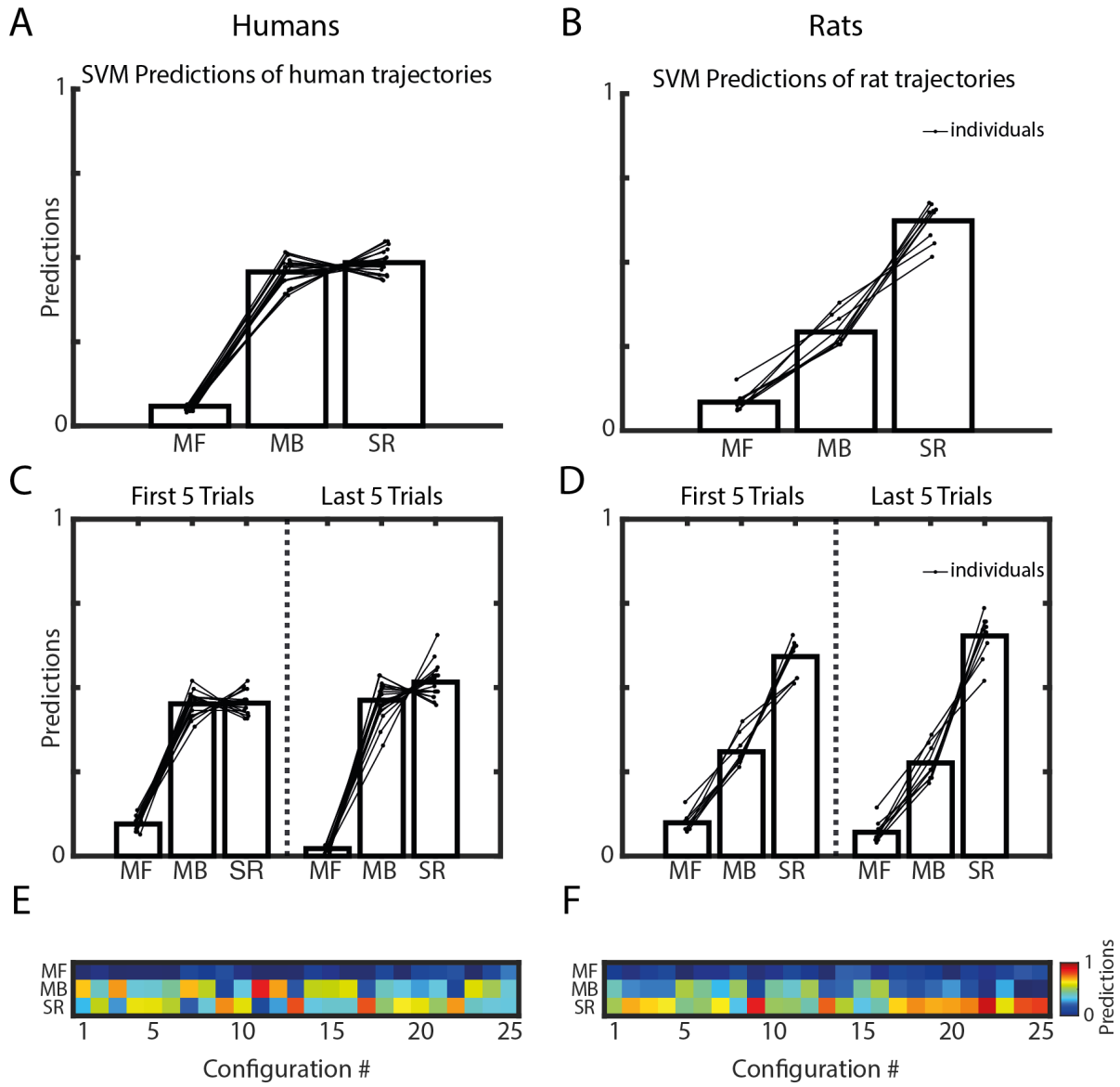


<u>Figure 5</u>: Trajectory similarity analysis. Analysis reveals that rat trajectories are most similar to a successor representation agent, whilst human trajectories are similar to both successor representation and model-based agents. (A) The shape of each trajectory was quantified using three metrics: the average diffusivity (B) and average angular diffusivity (C) provide a measure of the general dispersion and direction of a trajectory, whilst the tortuosity (D) gives a description of larger scale complexity. (E) Using these metrics to quantify trajectories for every start position on each maze, agent behaviour can be seen to form clusters (shown here via t-sne), where each point represents an individual human, rat or agent. Calculating the Mahalanobis distance between clusters reveals the human (F) and rat (G) behaviour is most similar to a successor representation agent.

Support vector machine classification identifies agent patterns in human and rat data

Finally, to test if the differences in these measures of the trajectory are sufficiently robust to allow categorisation, we trained a support vector machine (SVM) classifier to accurately predict trajectories as either model-free, model-based or SR agents (SFig3). When the decoder was given data from the biological behaviour, the SVM classifier most frequently predicted those trajectories to be an SR agent for rats (Fig. 6AB; SR vs MB: $t(8)=9.8$, $p<.001$ SR vs MF: $t(8)=11.8$, $p<.001$). Interestingly, there was no discernible difference in SR and model-based agent predictions in humans until the latter half of trials (Fig6CD; first 5 trials $t(17)=0.14$, $p=.89$; last 5 trials $t(17)=2.1$, $p=.048$). Viewing how SVM predictions change from maze to maze again reveals noticeable variation (Fig 6EF), with configurations that have little overlap in policy with the preceding maze often leading to increased SVM predictions of the SR agent.

In summary, we used three approaches to compare RL agents to rats and humans: a likelihood analysis of rats and humans actions under different agents, the similarity of RL agents performance and trajectories to rats and humans ones, and an unsupervised categorisation of rat and human trajectories by a support vector machine (SVM) classifier. Our results show that both species match more closely the SR RL agents' performance patterns than MF or MB agents, and that there are features of rodent and human navigation that remain more similar than the simulated agents.



<u>Figure 6</u>: Support vector machine (SVM) classifies rat trajectories as SR agents and humans as SR or MB agents. A SVM classifier trained to decode model-free (MF), model-based (MB) and successor representation (SR) agents most often classified human trajectories (left column) as MB and SR agents (A; SR vs MB: t(17)=1.51, p=0.15), and rat trajectories (right column) as an SR agent (B; SR vs MB: t(8)=9.8, *p*<.001). Later trials on a configuration are more likely to be classified as an SR agent for both the human (C; SR vs MB prediction on last 5 trials of maze, t(17)=2.1, *p*=.048) and rat data (D; SR vs MB prediction on last 5 trials of maze, t(8)=11.1, *p*<.001). The level of these predictions varied from maze-to-maze (E-F) with a strong correlation between the humans and rats (rank correlation: ρ = .71, *p*<.001).

Discussion

To understand the underlying processes that support flexible navigation in rats and humans we compared their navigation performance with three classic instantiations of RL agents in a maze environment with barriers that periodically reconfigured across trials. Using a combination of likelihood, trajectory similarity and support vector machine classifier analyses, we find rats and humans produce similar trajectory patterns that most resemble successor representation RL agents. This was most evident for rats, while humans were also found to show some trajectory similarity to model-based RL agents. Our findings provide cross-species translational insights into spatial navigation behaviour within a framework that allows comparison to simulated agents. This approach enables the prediction of specific metrics to capture future behaviour and neural states linked to models. We discuss: i) how these results relate to past failures and success in studying detours and shortcuts in both species, ii) similarities and differences between rodents and humans, iii) consideration of our task and analysis framework, and iv) directions for future research.

Rats and humans demonstrate similar flexible navigation in our dynamic environment

Past research has suggested that rats do not always optimally adapt to selecting appropriate alternative routes when navigating (Grieves & Dudchenko, 2013). Similarly, humans can struggle to take optimal shortcuts when presented with them (Brown et al., 2020; Foo et al., 2005). Here rats and humans had to reach the correct goal from a set of 100 possible goal locations within a 45 second time limit. Since the maze transition structure was re-configured every 10 trials, achieving this was not trivial. Despite this, both species were able to reach the goal on the first trial of a new layout on many trials. In several cases we saw examples of routes near the optimal path on the first attempt for both species (e.g. Video S1&2). Moreover, the occupancy correlation between rats and humans was relatively high even from the first trial, and improved as performance increased across trials in a configuration. These results show that our Tartarus maze, with its visual access to landmarks, boundary geometry and canyon-styled barriers, provides a useful assay for navigation across two species.

On the basis that humans engage in more detailed elaborate planning compared to rats (Passingham, 2021), and that rats are perhaps less flexible in their route choice (Grieves & Dudchenko, 2013), it could have been predicted that rats would most match the model-free agents and humans the model-based agents. This was not what we observed. Intriguingly, on a range of measures both species match most the successor representation which provides a computationally efficient mechanism for adapting to changes in the structure of the maze. What makes our observations striking is that while rats and humans performed better than the RL agents overall, they were not a consistent match with the best performing agents: model-based RL. Rather it was important to account for the patterns in the trajectories taken by the agents to reveal that the choices made in the mazes rather provided a better match to the patterns of choice by successor representation agents. Rats showed a consistently better capacity to adapt to the mazes compared to the model-free agents which, predictably, struggled to reach the goal when the maze configurations changed. It seemed plausible that rats with poorer performance

than humans might fall into following a learned ‘habit’ - often equated with model-free RL. However the latter half of trajectories on a configuration, where such habits might be more likely to come into place, were increasingly dominated by SR-like behaviour.

An advantage of having a large open space is that it provides a better match to real-world complexity and allows for a greater variation in possible behaviours to match to models. A simple t-maze task by comparison is much more limited in such capacity.

Our work parallels other recent work exploring navigation in rats and humans in the same framework and using water-maze tasks (Samanta et al., 2021), where rather than exploring changing terrain, updating new goal locations was examined. By incorporating a comparison to RL agents we are able to understand how species differ in relation to models and thus the potential mechanisms driving the behaviour - here with evidence favouring a successor representation instantiation of RL agents.

Our work also adds to findings that the SR provides an account of hippocampal representations observed in rats and humans(Brunec & Momennejad, 2019; de Cothi & Barry, 2020; Garvert et al., 2017; Sosa & Giocomo, 2021; Stachenfeld et al., 2017; but see Duvelle et al., 2021) by showing it often fits their spatial navigation behaviour in dynamic environments. We found less evidence for model-based agent behaviour in the rodent trajectories relative to the human behaviour, which may reflect less successful planning mechanisms or less propensity to use them. However, we did observe a high degree of overlap between the human and rat spatial behaviour relative to the artificial agents, suggesting both mammalian species may have evolved similar mechanisms for navigating in dynamic environments.

Comparison to previous experiments comparing behaviour to RL agents

Previous experiments comparing human behaviour to RL agents have generally focused on a dichotomy of model-free and model-based behaviours in small conceptual state spaces with 2-step transitions. These have found evidence for model-based planning in humans (Daw et al., 2011; Vikbladh et al., 2019; Wunderlich et al., 2012). By using a much larger state space with potential for recursive transitions (i.e. loops leading back to the same state), we extend this approach into a more complex and naturalistic framework. Our findings extend recent evidence that the actions of human participants are best explained by a combination of SR and model-based behaviours (Momennejad et al., 2017). Since the SR encodes the environment’s transition structure, it is itself a model that can be leveraged for intuitive planning (Baram et al., 2018) or more explicit planning procedures such as a tree search - which may also partially explain the trajectories observed in during hippocampal replay (Mattar & Daw, 2018; Momennejad, 2020; Momennejad et al., 2018; Pfeiffer & Foster, 2013).

Previous studies that have explored human navigation in VR environments with RL agents have reported a mix of model-based and model-free responses(Anggraini et al., 2018; Simon & Daw, 2011). However, these past studies did not compare with a successor representation, nor were the trajectories examined in relation metrics such as the

diffusivity to understand the match to the different RL agents. It will be useful to explore in future research the neural dynamics of navigating in the Tartarus maze and compare the responses to RL agents. See Future directions below for discussion.

How does the Tartarus Maze relate to other maze environments in helping understand navigation?

Prior lab based studies examining rodent navigation have used either tracks or open field environments(Poulter et al., 2018). While open field environments place more demands on self-localisation and vector-based navigation (Morris, 1984), mazes with tracks are helpful to examine the effect of blocked paths on behaviour (Grieves & Dudchenko, 2013; Tolman & Honzik, 1930; Alverhne et al., 2011). Instead, our open field maze task requires adaptation to changing obstacles and self-localisation. The recent development of the honey-comb maze (Wood et al 2018) provides a parallel approach to self-localisation and obstructed paths to goals, where rats sequentially navigate to a goal by selecting which maze module to step onto next. Such an approach allows the assessment of choice options at different time points. By contrast our task allows the comparison of RL agents with continual travel to goal locations. Rats adapted to the changes remarkably quickly, as shown by the failure of the model-free learner to match the behaviour. Rats rarely stopped to explore the regions of the change in the maze, which is more common in walled environments (Spiers et al., 2015). Prior human navigation studies exploring flexible behaviour tended to involve more complicated VR environments that would likely be too demanding for rodents to learn (A.-H. Javadi, Patai, Marin-Garcia, Margois, et al., 2019; A.-H. Javadi, Patai, Marin-Garcia, Margolis, et al., 2019) . Moreover, our current environment and task provides an ideal setting to explore single unit dynamics, optogenetic disruption in rodents, in parallel with fMRI or MEG in humans.

Differences between RL agents

There was a noticeable improvement in successful goal-reaching between the biological behaviour and RL agents. The learning efficiency of RL agents can be improved using offline replay of randomly sampled past experiences (Russek et al., 2017; Sutton, 1991). These replays are typically implemented between agent timesteps, and the manner in which they are sampled can further accelerate learning by prioritising the most useful learning experiences to replay (Mattar & Daw, 2018). Prioritised replay also has strong parallels to the phenomenon of hippocampal replay (Liu et al., 2019; Pfeiffer & Foster, 2013). However, we did not implement agent replay in this study in order to keep the value representations and consequently the likelihoods of agents deterministic enabling fair model comparison. An alternative way to improve the goal-reaching of agents could be through improving their exploration policy. The agents simulated here relied on an epsilon-greedy policy through which exploration is purely driven by chance. However, methods that include curiosity (Still & Precup, 2012) or uncertainty in the value function (Geerts et al., 2019; Gershman, 2015) could be used to guide more efficient exploration of a new maze and consequently lead to faster learning.

Considerations in the analysis approach we adopted

The ubiquitous importance of spatial navigation across animals makes it an excellent candidate framework for studying the brain and behaviour across species. However, comparing trajectories that can vary in length and ultimate destination is not a trivial problem to solve. The method used in this study involved characterising the trajectories with metrics that capture their general shape and complexity. Consequently, similarity was calculated with respect to this characterisation. Since the measures will inevitably have some degree of correlation, it is important to use a dissimilarity measure that takes this covariance into account. The Mahalanobis scales distances across axes by their covariance, and has also been successfully used as a dissimilarity measure for patterns of neural activity (Kriegeskorte et al., 2008). An interesting avenue for developing this would be to extract the trajectory measures in real-time and produce a continually updated, real-time prediction of the match to RL agents during navigation.

Future directions

Our paradigm allows for the capacity to test the prediction that the hippocampal activity matches model predications that the hippocampus implements SR for navigation. Parallel recording of human and rodent neural data in the same task and comparison to RL agents allow for the predictions about neural dynamics at different events in the navigation process to be examined with different methods and integrated for a better understanding (Gahnstrom & Spiers, 2020). Past studies have shown that dorsal striatal responses scale with the amount of change in the path to the goal after a change in the terrain (Howard et al., 2014; A.-H. Javadi, Patai, Marin-Garcia, Margolis, et al., 2019) (Howard et al, 2014; Javadi et al., 2019) and that anterior cingulate activity enhances when participants choose to back-track (A.-H. Javadi, Patai, Marin-Garcia, Margolis, et al., 2019). Both these behaviours (detour choice and back-tracking) might be predictable from a suitable model and RL would provide the capacity to predict the level or response elicited. For example it may be possible to predict the magnitude of the anterior cingulate response from a measure of how surprising the model finds the change in the maze layout. An advantage of adopting a similar framework for testing is that it allows future neural recording studies to integrate over different methods to study neural responses in rodents and humans (Barron et al., 2020, 2021; Samanta et al., 2021). Such an approach would also aid the translational neuroscience assessment of interventions for spatial navigation, which is affected early in alzheimers' disease.

While both species are overall successful at performing the task. This makes it possible to use such a task to examine the neural codes for information salient to the task (e.g. path to the goal) and also understand the underlying mechanisms by which the animals navigated (see Patai & Spiers, 2021).

Conclusion

In summary, we found that rats and humans both display behaviour most similar to a successor representation RL agent, with humans also showing some model-based planning. Future work exploring single unit recording or disruption to neural activity may

be useful in revealing how distance to the goal may be coded, as past studies have failed to dissociate path and Euclidean distance. Moreover it will be useful to examine how neural activity in humans and rodents relates to the parameters from RL agents with behaviour adjusted to match the humans and rats.

Materials and Methods

General methods

Navigation was tested in a large square environment with a fixed hidden goal location and a prominent directional black wall cue in one direction (Fig. 1, Video S1 and S2). The maze was divided in a 10x10 grid of moveable sections that could either be removed, leaving impassable gaps to force detour taking, or added, creating shortcuts. During training, all maze modules were present. Rats, humans and RL agents were trained to reach the goal within a 45s time limit, (Fig 1A). During the testing phase of the experiment, maze modules were removed to block the direct route to the goal (Fig 1B). Humans (n=18), rats (n=9) and agents were tested on the same sequence of 25 maze configurations each with 10 trials in which a set of defined starting locations were selected to optimally probe navigation (Fig 1C). These maze configurations were generated from a pilot testing with 9 rats and the maze sequence chosen maximized the differences in the layouts between trials. The starting positions on each maze gradually increased in the required tortuosity (path distance / Euclidean distance) of the shortest path to the goal to test complex trajectories whilst keeping the rodents motivated.

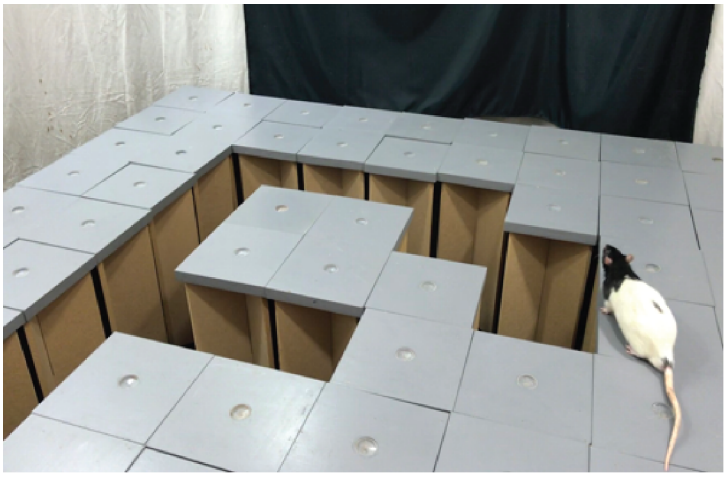
Upon reaching the goal module, rats and humans had to wait 5s to receive their reward. Human participants were rewarded with a financial bonus and rats received chocolate milk delivered in a well (supp. fig 1). In order to better match the visual acuity and depth perception between rodents and humans (Heffner & Heffner, 1992), a thick virtual fog lined the floor of the maze enabling them to only see adjacent maze modules and the distal black wall cue (supp. fig 2; Video S3). Modules were made visually indistinct to avoid humans counting them when traversing the space. Human participants were informed that reward was hidden in the environment and that their task was to maximise their financial return as quickly and efficiently as possible. The human and rat trajectories were discretised into the underlying 10x10 modular grid (Fig 2A-B) in order to facilitate comparison between each other and the RL agents.

In all versions of the experiment, the environment (raised off the floor) consisted of a 10x10 grid of maze modules. These modules could be removed from the grid in order to form impassable barriers in the environment. One of the modules was rewarded and thus was the location of the goal in the maze. Navigation was facilitated by a single distal cue consisting of a black curtain that spanned the majority of one side of the maze. The goal was kept in the same position with respect to this distal cue throughout all versions of the task. All participants, rats and learning agents were initially trained to navigate to the goal module on the open maze, without any maze modules removed. Once trained, they were all put through the same sequence of 25 mazes, with the same sequence of starting locations on each maze

Rodent methods

Nine adult male Lister Hooded rats were handled daily (at start of training: 10-20 weeks old, 350-400 g) and housed communally in groups of three. All rats were subjected to a reverse light-dark cycle (11:11 light:dark, with 1 hour x2 simulated dawn/dusk) and were on food-restriction sufficient to maintain 90% of free-feeding weight, with ad libitum access to water. The free-feeding weight was continuously adjusted according to a calculated growth curve for Lister Hooded Rats (Clemens et al., 2014). Six rats were naive, while three rats had previously been trained for 2-3 weeks in a shortcut navigation task for a different maze setup. The procedures were conducted according to UCL ethical guidelines and licensed by the UK Home Office subject to the restrictions and provisions contained in the Animals Scientific Procedures Act of 1986.

All procedures were conducted during the animals' dark period. The experiment was carried out in a custom-made modular 2x2m square maze composed of 100 identical square platform tiles elevated 50cm above the ground (supp. fig 1). The maze was constructed from Medium Density Fibrewood, with the platforms painted in grey. Each platform contained a plastic well (32mm diameter) at its centre, which could be attached to a polymeric tubing system installed beneath the maze. This tubing allowed the experimenter to reward the rat at the goal module filling the well with chocolate milk (0.1 ml). Importantly, all modules in the rodent maze were identical in appearance and construction with chocolate milk rubbed into the well of non-goal modules to lower reliance on olfactory navigational cues. The maze was surrounded on all sides by a white curtain, with a black sheet overlaid on one side to provide a single extra-maze cue. To ensure that no other cue could be used by the animal (uncontrolled room cues, olfactory traces on the maze) the black sheet was rotated 90° clockwise between sessions. The goal module was always in the same position with respect to this cue. Moreover, the experimenter stayed next to the maze inside the curtained area throughout all sessions, his positions relative to the goal were randomised.



Supplementary Figure 1: The maze environment used for the rat experiment. The environment consists of 100 removable mazes modules with a black curtain over one of

the surrounding edges to provide a single extra-maze cue. Reward can be dispensed at the goal module by filling the well with chocolate milk via polymeric tubing beneath the maze.

Familiarisation. During the first day, the rats received a small amount (0.1ml per rat) of chocolate milk in the home cage to decrease neophobia on the maze. For the subsequent two days, each rat underwent two 15 minute maze familiarisation sessions, in which the rat was placed at the centre of the maze and would forage for pieces of chocolate cereal (Weetos) scattered throughout the maze. More cereal was concentrated in the centre to encourage the animal to be comfortable in the middle of the maze. .

Training.. Training consisted of two stages, rats were given 2 training sessions per day. In each training trial the rat had 45s to find the goal module.

Stage 1. The goal well was filled with 0.1ml of chocolate milk and the rats were initially placed on the modules adjacent to the goal, facing the goal. If the rat made two consecutive direct runs to the goal (without exploration of other parts of the maze), the next trial begun one module further away from the goal. Conversely, if the rat failed two consecutive training trials, the next trial begun one module closer to the goal until the rat was back at the goal-adjacent modules. On day 1, this procedure was continued until 15 min had elapsed.On the following days, the number of trials was fixed to 16. This procedure was followed every day until the rat was able to make direct runs from the far edges of the maze.

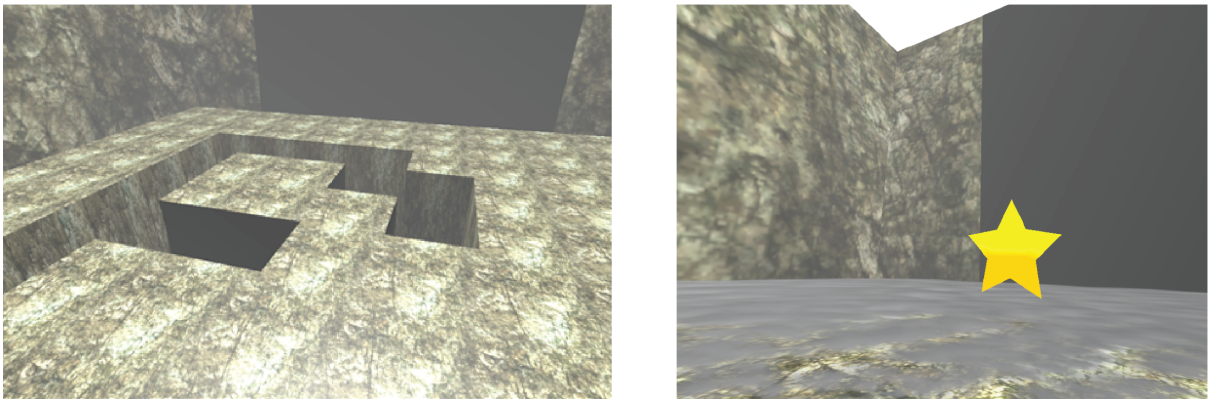
Stage 2 was also similar to stage1 but a delay in the release of chocolate milk was introduced. This delay started at 1s and was gradually increased until the rat could wait at the goal location for 5s before the chocolate milk was released. Furthermore, the rat's starting position and orientation were randomised. The number of daily trials could be increased up to 25. This procedure was followed until the rats were able to successfully navigate directly to the goal and on at least 90% of trials. The training phase took on average 24 sessions.

Tartarus Maze testing. - Rats were run on the 25 maze configurations. For each maze configuration, rats were given 10 trials, starting from with the starting positions indicated in figure 1. Trials were 45s long and rats were required to navigate to the goal within this time and wait for 5s in order to receive the reward (0.1ml of chocolate milk). If the rat failed to reach the goal, it moved onto the next trial. The rats would usually complete 3 configurations per day. At the beginning of each day, rats were given a brief reminder session that consisted of 5 trials from phase 2 of the training phase.

Human methods

For the human version of the task, 18 healthy participants (9 female; aged = 24.6 ± 5.9, mean ± sd) were recruited from the UCL Psychology Subject Pool and trained to navigate to an unmarked goal in a virtual arena of approximately the same relative proportion as

for the rats. All participants gave written consent to participate in the study in accordance with the UCL Research Ethics Committee. Participants were reimbursed for their time as well as a bonus of up to £25 for good performance in the testing phase. Participants experienced the virtual environment via a HTC Vive virtual reality headset whilst sat on a swivel chair. They were able to adjust movement speed using the HTC Vive controller and movement direction was controlled by the participant's orientation on the chair. Upon successful navigation to the goal module, participants were informed of their financial reward along with the presence of a revolving gold star (supp. fig 2) at the goal location. In accordance with the rodent experiment, navigation was aided by the presence of a black distal cue that took up the majority of one of the walls. Goal location, maze configurations and starting positions were all defined with respect to this distal cue and were identical to the rodent experiment. Importantly, a fog lined the floor (supp. fig 2, Video S2) of the maze to prevent the participants from understanding what maze modules were missing until they were at adjacent locations. This also provided a better match to visual information available to the rats - which are known to have less visual acuity and binocular depth perception (Heffner & Heffner, 1992). Seamless textures were applied to the floor and walls of the virtual environment, and these were rotated every 10 trials to prevent them from being used as extraneous cues for navigation.



Supplementary Figure 2: The virtual environment used for the human experiment. The environment had the same proportions as the rat environment and consisted of 100 removable mazes modules with a black curtain over one of the surrounding edges to provide a single extra-maze cue. A seamless texture was applied to the maze modules and walls and a fog lined the floor of the maze (see right image) to ensure humans had to rely on spatial memory to understand the maze structure. Reward was indicated by a gold star that would appear at the goal module when the participant successfully navigated to it.

The experiment took place over four sessions on four consecutive days. The majority of the first session was usually spent training the participants to navigate to the goal module. To accelerate this learning process, the participants were initially able see a revolving gold star in the goal location. As they progressed through the training session the star became increasingly transparent until invisible, with the star only appearing again upon successful navigation to the goal module. Along with the decreasing visibility of the goal, the participants' starting positions were moved progressively further from the goal in a similar manner to the rat training phase. All training and testing trials were 45s in length. Training was terminated when the participants were able to navigate to the hidden goal on at least

80% of trials after being randomly placed at the far edges of the environment. Mean time to complete this training was 41 ± 21 minutes. In order to make the participants' experience similar to that of the rodents, they were not given any explicit information about the nature of the task - only that financial reward was hidden in the environment in the form of a gold star and their task was to maximise their financial return as quickly and efficiently as possible.

The testing took place over the remaining sessions and on average lasted 125 ± 25 minutes, with participants encouraged to take short breaks every 10-20 trials to reduce virtual reality sickness. At the beginning of each testing session, participants completed a short reminder task, which consisted of 5 trials from the end of the training phase.

Reinforcement learner simulations

The reinforcement learning agents were implemented in a 10x10 grid world. At the beginning of the experiment, all agents were endowed with the optimal policy on the open maze to simulate the training phase undertaken by rats and humans. They were then run consecutively on the 25 maze configuration, carrying over all value and model representations between trials. All agent behaviour was simulated using the maximum likelihood parameters fit to the rat or human data, along with an ε-greedy policy where ε = 0.1. This means the agents choose the greedy action 90% of the time and a random action 10% of the time (in order to manage the exploration-exploitation tradeoff). Due to the behavioural variance introduced by this policy, each algorithm was implemented 1000 times to produce the distribution of behaviour used for comparison with the rats and humans.

The model-free agents were simulated by implementing equation 5 at every time step to update the action-value function. The SR agents were simulated by implementing equations 7 and 8 at every time step to update the SR.
The model-based agent updated its model at every time step by observing the states adjacent to its current state and implementing the update in equation 6.
It would then sample a map from the model and use it to plan the shortest route to the goal from its current position (shortest routes were calculated using the A-star search algorithm (Hart et al., 1968). In the event of multiple equally short routes to the goal, their respective actions were sampled with equal probability.

Model-free agent

The model-free agent uses a Q-learning algorithm (Watkins & Dayan, 1992) to experientially update the state-value function Q (equation 2). Specifically, after taking action a_t in state s_t and transitioning to state s_{t+1} where it receives reward r_t, the agent will update the state-action value according to:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\Big[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\Big] \tag{5}$$

where α is the learning rate and γ is the discount factor of the value function. Under a greedy policy, the model-free agent at decision time will choose an action a in state s with the highest state-action value $Q(s, a)$.

Model-based agent

The model-based agent is provided with an internal 10x10 binary grid representation of which maze modules are present or not in the environment. Every state s in the agent's model χ corresponds a module in the maze (see Fig 1A-B); as it transitions through the environment, it updates the internal model at every timestep according to the adjacent states s'.

$$\chi(s') \leftarrow \begin{cases} 1 & \text{if module } s' \text{ is present} \\ 0 & \text{if module } s' \text{ is missing} \end{cases} \tag{6}$$

At decision time, the model-based agent uses its model χ and to plan the shortest route to the goal from each possible next state. Shortest routes were calculated using an A* tree search algorithm (Hart et al., 1968), and that the value of each prospective future state was computed via equation 1.

Successor representation agent

The SR agent uses temporal-difference learning to update the successor matrix M

$$M(s, :) \leftarrow M(s, :) + \alpha[\mathbb{1}_s + M(s', :) - M(s, :)] \tag{7}$$

which can be combined with the state-rewards R at decision time to compute the value of prospective future states via equation 4.

$$R(s) = \begin{cases} 1 & \text{if } s \text{ is the goal} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

All agents implement an ε-greedy policy with $\varepsilon = 0.1$, meaning at each timestep they will choose the highest value next state with probability 0.9, or a random next state with probability 0.1. All internal representations used by the SR agent and both other RL agents were carried over onto the first trial of the subsequent maze.

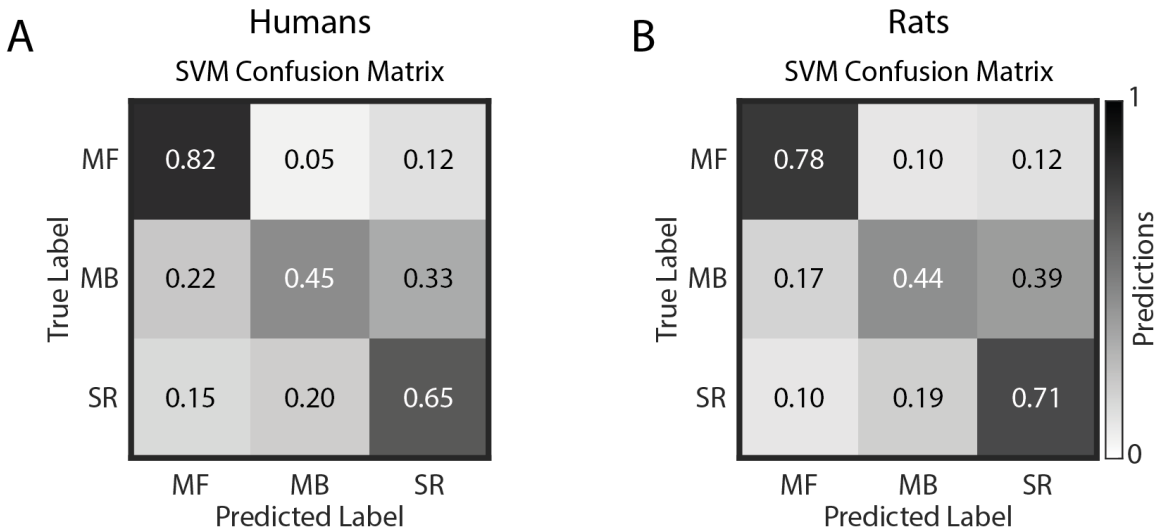
Quantification and Statistical Analysis

Optimal paths were calculated using the A* tree search algorithm (Hart et al., 1968) in the 10x10 grid state space, with path length measured in terms of state visitations. Occupancy correlations were calculated using the Pearson correlation between the proportion of time spent in each state of the 10x10 grid state space.

Likelihoods were calculated by inputting individual human/rat state trajectories to a given RL agent and calculating the internal value estimates of the available state transitions conditional on the human/rat's past trajectories. Finally, these value estimates were used in a softmax function to calculate at each time point, the probability that the agent would take each of the available actions conditioned on the human/rat's past. Maximum likelihood parameters were estimated using MATLAB's fmincon function to minimise the negative log-likelihood.

Mahalanobis distances were calculated per trial using MATLAB's pdist2 function on the trajectory metrics for the humans, rats, model-free, model-based and successor representation agents.

Support vector machine (SVM) classifiers were trained and assessed using 5-fold cross validation to decode RL agent identities simulated with the human (SFig3a) and rat (SFig3b) maximum likelihood parameters. Individual SVM's were fit per trial using MATLAB's fitcecoc function and a one vs one coding scheme.



Supplementary Figure 3: Confusion matrices for the support vector machine (SVM) classifier. The SVM classifier was able to accurately decode model-free (MF), model-based (MB) and successor representation (SR) agent identity based on a 5-fold cross-validation and the maximum likelihood parameters estimated from the human (A) and rat (B) behaviour.

Acknowledgments

We would like to thank Célia Lacaux and Charles Middleton for their help building the maze.

CB is funded by Wellcome SRF (212281/Z/18/Z). WdC was funded by an EPSRC CASE studentship. NN and HJS received support from the European Union's Horizon 2020 Framework Programme for Research under Marie Sklodowska-Curie ITN (EU-M-GATE 765549). HJS received funding from Deepmind Ltd.

References

Alvernhe, A., Cauter, T. V., Save, E., & Poucet, B. (2008). Different CA1 and CA3 Representations of Novel Routes in a Shortcut Situation. *Journal of Neuroscience*, *28*(29), 7324–7333. https://doi.org/10.1523/JNEUROSCI.1909-08.2008

Andersen, P., Morris, R., Amaral, D., Bliss, T., & O'Keefe, J. (2006). *The hippocampus book*. Oxford university press.

Anggraini, D., Glasauer, S., & Wunderlich, K. (2018). Neural signatures of reinforcement learning correlate with strategy adoption during spatial navigation. *Scientific Reports*, *8*(1), 10110. https://doi.org/10.1038/s41598-018-28241-z

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., Wayne, G., Soyer, H., Viola, F., Zhang, B., Goroshin, R., Rabinowitz, N., Pascanu, R., Beattie, C., Petersen, S., … Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429–433. https://doi.org/10.1038/s41586-018-0102-6

Baram, A. B., Muller, T. H., Whittington, J. C. R., & Behrens, T. E. J. (2018). Intuitive planning: Global navigation through cognitive maps based on grid-like codes. *BioRxiv*, 421461. https://doi.org/10.1101/421461

Barron, H. C., Mars, R. B., Dupret, D., Lerch, J. P., & Sampaio-Baptista, C. (2021). Cross-species neuroscience: Closing the explanatory gap. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1815), 20190633. https://doi.org/10.1098/rstb.2019.0633

Barron, H. C., Reeve, H. M., Koolschijn, R. S., Perestenko, P. V., Shpektor, A., Nili, H.,

Rothaermel, R., Campo-Urriza, N., O'Reilly, J. X., Bannerman, D. M., Behrens, T. E. J., & Dupret, D. (2020). Neuronal Computation Underlying Inferential Reasoning in Humans and Mice. *Cell*, *183*(1), 228-243.e21. https://doi.org/10.1016/j.cell.2020.08.035

Bellmund, J. L. S., de Cothi, W., Ruiter, T. A., Nau, M., Barry, C., & Doeller, C. F. (2019). Deforming the metric of cognitive maps distorts memory. *Nature Human Behaviour*, 1–12.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep Reinforcement Learning and Its Neuroscientific Implications. In *Neuron* (Vol. 107). Cell Press. https://doi.org/10.1016/j.neuron.2020.06.014

Brown, T. I., Gagnon, S. A., & Wagner, A. D. (2020). Stress Disrupts Human Hippocampal-Prefrontal Function during Prospective Spatial Navigation and Hinders Flexible Behavior. *Current Biology*, *30*(10), 1821-1833.e8. https://doi.org/10.1016/j.cub.2020.03.006

Brunec, I. K., & Momennejad, I. (2019). Predictive Representations in Hippocampal and Prefrontal Hierarchies. *BioRxiv*, 786434. https://doi.org/10.1101/786434

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624. https://doi.org/10.1162/neco.1993.5.4.613

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. In *Cognitive, Affective and Behavioral Neuroscience* (Vol. 8). Springer. https://doi.org/10.3758/CABN.8.4.429

de Cothi, W., & Barry, C. (2020). Neurobiological successor features for spatial navigation. *Hippocampus*, *30*(12), 1347–1355. https://doi.org/10.1002/hipo.23246

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron*, *80*, 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Duvelle, É., Grieves, R. M., Liu, A., Jedidi-Ayoub, S., Holeniewska, J., Harris, A., Nyberg, N., Donnarumma, F., Lefort, J. M., Jeffery, K. J., Summerfield, C., Pezzulo, G., & Spiers, H. J. (2021). Hippocampal place cells encode global location but not connectivity in a complex space. *Current Biology*, *31*(6), 1221-1233.e9. https://doi.org/10.1016/j.cub.2021.01.005

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, *425*(6954), 184–188.

Ekstrom, A. D., & Ranganath, C. (2018). Space, time, and episodic memory: The hippocampus is all over the cognitive map. *Hippocampus*, *28*(9), 680–687. https://doi.org/10.1002/hipo.22750

Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience*, *20*(11), 1504–1513. https://doi.org/10.1038/nn.4656

Foo, P., Warren, W. H., Duchon, A., & Tarr, M. J. (2005). Do humans integrate routes into a cognitive map? Map- Versus landmark-based navigation of novel shortcuts. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(2), 195–215. https://doi.org/10.1037/0278-7393.31.2.195

Gahnstrom, C. J., & Spiers, H. J. (2020). Striatal and hippocampal contributions to flexible navigation in rats and humans. *Brain and Neuroscience Advances*, *4*, 2398212820979772. https://doi.org/10.1177/2398212820979772

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *ELife*, *6*, 1–20. https://doi.org/10.7554/eLife.17086

Geerts, J., Stachenfeld, K., & Burgess, N. (2019). Probabilistic Successor Representations with Kalman Temporal Differences. *Conference on Cognitive Computational Neuroscience*. https://doi.org/10.32470/ccn.2019.1323-0

Gentry, G., Brown, W. L., & Kaplan, S. J. (1947). An experimental analysis of the spatial location hypothesis in learning. *Journal of Comparative and Physiological Psychology*, *40*(5), 309.

Gershman, S. J. (2015). A Unifying Probabilistic View of Associative Learning. *PLoS Computational Biology*, *11*(11), 1–20. https://doi.org/10.1371/journal.pcbi.1004567

Gershman, S. J. (2018). The Successor Representation: Its Computational Logic and Neural Substrates. *Journal of Neuroscience*, *38*(33), 7193–7200. https://doi.org/10.1523/JNEUROSCI.0151-18.2018

Grieves, R. M., & Dudchenko, P. A. (2013). Cognitive maps and spatial inference in animals: Rats fail to take a novel shortcut, but can take a previously experienced one. *Learning and Motivation*, *44*(2), 81–92. https://doi.org/10.1016/j.lmot.2012.08.001

Grieves, R. M., & Jeffery, K. J. (2017). The representation of space in the brain. *Behavioural Processes*. https://doi.org/10.1016/j.beproc.2016.12.012

Howard, L. R., Javadi, A. H., Yu, Y., Mill, R. D., Morrison, L. C., Knight, R., Loftus, M. M., Staskute, L., & Spiers, H. J. (2014). The hippocampus and entorhinal cortex encode the path and euclidean distances to goals during navigation. *Current Biology*, *24*(12), 1331–1340. https://doi.org/10.1016/j.cub.2014.05.001

Javadi, A. H., Emo, B., Howard, L. R., Zisch, F. E., Yu, Y., Knight, R., Pinelo Silva, J., & Spiers, H. J. (2017). Hippocampal and prefrontal processing of network topology to

simulate the future. *Nature Communications*, *8*(1), 1–11. https://doi.org/10.1038/ncomms14652

Javadi, A.-H., Patai, E. Z., Marin-Garcia, E., Margois, A., Tan, H.-R. M., Kumaran, D., Nardini, M., Penny, W., Duzel, E., Dayan, P., & Spiers, H. J. (2019). Backtracking during navigation is correlated with enhanced anterior cingulate activity and suppression of alpha oscillations and the 'default-mode' network. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1908), 20191016. https://doi.org/10.1098/rspb.2019.1016

Javadi, A.-H., Patai, E. Z., Marin-Garcia, E., Margolis, A., Tan, H.-R. M., Kumaran, D., Nardini, M., Penny, W., Duzel, E., Dayan, P., & Spiers, H. J. (2019). Prefrontal Dynamics Associated with Efficient Detours and Shortcuts: A Combined Functional Magnetic Resonance Imaging and Magnetoencenphalography Study. *Journal of Cognitive Neuroscience*, *31*(8), 1227–1247. https://doi.org/10.1162/jocn_a_01414

Laczó, J., Andel, R., Vyhnalek, M., Vlcek, K., Magerova, H., Varjassyova, A., Nedelska, Z., Gazova, I., Bojar, M., Sheardova, K., & Hort, J. (2012). From Morris Water Maze to Computer Tests in the Prediction of Alzheimer's Disease. *Neurodegenerative Diseases*, *10*(1–4), 153–157. https://doi.org/10.1159/000333121

Lee, J. J., & Keramati, M. (2017). Flexibility to contingency changes distinguishes habitual and goal-directed strategies in humans. *PLoS Computational Biology*, *13*(9). https://doi.org/10.1371/journal.pcbi.1005753

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human Replay Spontaneously Reorganizes Experience. *Cell*, *178*(3), 640-652.e14. https://doi.org/10.1016/j.cell.2019.06.012

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*(11), 1609–1617.

https://doi.org/10.1038/s41593-018-0232-z

Miller, K. J., Botvinick, M. M., & Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, *20*(9), 1269–1276. https://doi.org/10.1038/nn.4613

Momennejad, I. (2020). Learning Structures: Predictive Representations, Replay, and Generalization. In *Current Opinion in Behavioral Sciences* (Vol. 32). Elsevier Ltd. https://doi.org/10.1016/j.cobeha.2020.02.017

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *ELife*, *7*, 1–25. https://doi.org/10.7554/eLife.32548

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692. https://doi.org/10.1038/s41562-017-0180-8

Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods*, *11*(1), 47–60. https://doi.org/10.1016/0165-0270(84)90007-4

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

Passingham, R. E. (2021). *Understanding the prefrontal cortex: Selective advantage, connectivity, and neural operations* (Vol. 53). Oxford University Press.

Patai, E. Z., Javadi, A. H., Ozubko, J. D., O'Callaghan, A., Ji, S., Robin, J., Grady, C., Winocur, G., Rosenbaum, R. S., Moscovitch, M., & Spiers, H. J. (2019). Hippocampal and Retrosplenial Goal Distance Coding after Long-term Consolidation of a Real-World Environment. *Cerebral Cortex*, *29*(6), 2748–2758. https://doi.org/10.1093/cercor/bhz044

Patai, E. Z., & Spiers, H. J. (2021). The Versatile Wayfinder: Prefrontal Contributions to Spatial Navigation. *Trends in Cognitive Sciences*, *25*(6), 520–533. https://doi.org/10.1016/j.tics.2021.02.010

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, *497*. https://doi.org/10.1038/nature12112

Poulter, S., Hartley, T., & Lever, C. (2018). The Neurobiology of Mammalian Navigation. *Current Biology*, *28*(17), R1023–R1042. https://doi.org/10.1016/j.cub.2018.05.050

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, *13*(9). https://doi.org/10.1371/journal.pcbi.1005768

Samanta, A., van Rongen, L. S., Rossato, J. I., Jacobse, J., Schoenfeld, R., & Genzel, L. (2021). Sleep Leads to Brain-Wide Neural Changes Independent of Allocentric and Egocentric Spatial Training in Humans and Rats. *Cerebral Cortex*, *bhab135*. https://doi.org/10.1093/cercor/bhab135

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Simon, D. A., & Daw, N. D. (2011). Neural Correlates of Forward Planning in a Spatial Decision Task in Humans. *Journal of Neuroscience*, *31*(14), 5526–5539. https://doi.org/10.1523/JNEUROSCI.4647-10.2011

Sosa, M., & Giocomo, L. M. (2021). Navigating for reward. *Nature Reviews Neuroscience*,

22(8), 472–487. https://doi.org/10.1038/s41583-021-00479-z

Spiers, H. J., & Barry, C. (2015). Neural systems supporting navigation. *Current Opinion in Behavioral Sciences*, *1*, 47–55. https://doi.org/10.1016/j.cobeha.2014.08.005

Spiers, H. J., & Gilbert, S. J. (2015). Solving the detour problem in navigation: A model of prefrontal and hippocampal interactions. *Frontiers in Human Neuroscience*, *9*(March), 1–15. https://doi.org/10.3389/fnhum.2015.00125

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*. https://doi.org/10.1038/nn.4650

Still, S., & Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, *131*(3), 139–148. https://doi.org/10.1007/s12064-011-0142-z

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44. https://doi.org/10.1007/bf00115009

Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, *2*(4), 160–163. https://doi.org/10.1145/122344.122377

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

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

Tolman, E. C., & Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology*.

Tolman, E. C., Ritchie, B. F., & Kalish, D. (1946). Studies in spatial learning. I. Orientation and the short-cut. *Journal of Experimental Psychology*, *36*(1), 13.

Uylings, H. B. M., Groenewegen, H. J., & Kolb, B. (2003). Do rats have a prefrontal cortex? In *Behavioural Brain Research* (Vol. 146). Elsevier. https://doi.org/10.1016/j.bbr.2003.09.028

Vikbladh, O. M., Meager, M. R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., Burgess, N., & Daw, N. D. (2019). Hippocampal Contributions to Model-Based Planning and Spatial Memory. *Neuron*, *102*(3), 683-693.e4. https://doi.org/10.1016/j.neuron.2019.02.014

Wallace, D. J., Greenberg, D. S., Sawinski, J., Rulla, S., Notaro, G., & Kerr, J. N. D. (2013). Rats maintain an overhead binocular field at the expense of constant fusion. *Nature*, *498*(7452), 65–69. https://doi.org/10.1038/nature12153

Watkins, C. J. C. H., & Dayan, P. (1992). Q-Learning. *Machine Learning*, *8*(3–4), 279–292. https://doi.org/10.1007/BF00992698

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine Enhances Model-Based over Model-Free Choice Behavior. *Neuron*, *75*, 418–424. https://doi.org/10.1016/j.neuron.2012.03.042

Xu, J., Evensmoen, H. R., Lehn, H., Pintzka, C. W. S., & H\aaberg, A. K. (2010). Persistent posterior and transient anterior medial temporal lobe activity during navigation. *NeuroImage*, *52*(4), 1654–1666. https://doi.org/10.1016/j.neuroimage.2010.05.074