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 Lefort[1], Lydia Fletcher[2], Charlotte Newton[4], Sophie Renaudineau[2], Daniel Bendor[2], Roddy Grieves[2], Éléonore Duvelle[2], Caswell Barry[1] & Hugo J. Spiers[2*]

1. Research 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

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

Abstract

Much of our understanding of navigation has come from the study of rats, humans and simulated artificial agents. To date little attempt has been made to integrate these approaches into a common framework to understand mechanisms that may be shared across mammals and the extent to which different instantiations of agents best capture mammalian navigation behaviour. Here, we report a comparison of rats, humans and reinforcement learning (RL) agents in a novel open-field navigation task ('Tartarus Maze') requiring dynamic adaptation (shortcuts and detours) to changing obstructions in the path to the goal. We find humans and rats are remarkably similar in patterns of choice in the task. The patterns in their choices, dwell maps and changes over time reveal that both species show the greatest similarity to RL agents utilising a predictive map: the successor representation. Humans also display trajectory features similar to a model-based RL agent. Our findings have implications for models seeking to explain mammalian navigation in dynamic environments and highlight the utility of modelling the behaviour of different species in the same frame-work in comparison to RL agents to uncover the potential mechanisms used for behaviour.

Key words: Artificial intelligence, Spatial Memory, Decision-making, Planning, Hippocampus

Introduction

For many animals knowing how to navigate to resources such as food and safety is a crucial part of everyday existence. In mammals it has been argued that the hippocampus supports navigation via the construction of a ‘cognitive map’ of the environment (O’Keefe & Nadel, 1978). Such an internal map allows the animal to infer novel spatial trajectories via shortcuts, plan future routes and efficiently adapt to changes in the environment by taking optimal detours (Epstein et al., 2017; Edward C Tolman, 1948).

Evidence from testing rodents and humans in real and virtual mazes has revealed an impressive ability for both species to take shortcuts and detours (Grieves & Dudchenko, 2013; Howard et al., 2014; Javadi et al., 2017; Javadi, Patai, Marin-Garcia, Margolis, et al., 2019; Patai et al., 2019; Spiers & Maguire, 2006; Xu et al., 2010), and to back-track when in error (Javadi, Patai, Marin-Garcia, Margois, et al., 2019). However there is wide variation in ability with both species also able to make considerable errors (Javadi, Patai, Marin-Garcia, Margolis, et al., 2019; Edward C Tolman, 1948). Rats often require multiple exposures to a set of paths in order to shift towards an optimal detour, and are far from optimal in selecting shortcuts if the paths presented are all novel (Grieves & Dudchenko, 2013). Humans too can make significant errors in judgment in judging the connectivity of spaces (Foo et al., 2005) with older humans often failing to adopt useful shortcuts (Harris & Wolbers, 2012). Consequently, developing a task to test flexible route-taking in both humans and rats is non-trivial.

Errors in path choices when adapting to changes in the environment have the potential to reveal the underlying mechanisms supporting flexible navigation (Gahnstrom and Spiers, 2020). They also may inform us of similarities and differences between species. Much research into navigation implicitly assumes that rats and humans fundamentally navigate in a similar way (Ekstrom & Ranganath, 2018; Epstein et al., 2017) and are used to support the integration of insights from both species (O’Keefe and Nadel, 1978; Spiers and Barry, 2015). However, very few experiments have sought to compare rats and humans on a directly homologous task. Given that humans and rats differ in both sensory processing (e.g. differences in binocular vision, whiskers, olfactory function) and prefrontal regions (Uylings et al., 2003) believed to support spatial planning (Spiers, 2008; Spiers & Gilbert, 2015), it is plausible that analyses of the navigational similarity may reveal important differences between these two species. Moreover, it is essential that such an approach examines the spatial trajectories in manner that can tease apart potential mechanisms underlying the behaviour (Gahnstrom and Spiers 2020).

One approach for teasing apart potential mechanisms underlying behaviour is the application of reinforcement learning models (Banino et al., 2018; Botvinick et al., 2020; Dayan & Daw, 2008; Ida 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 accumulated reward yielded from its expected future states , which is known as the value function V:

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

Here is a discount factor parameter 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 which is 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. Model-free methods are generally more simple and inexpensive to compute through a reliance on temporal-difference (TD) learning rules (Sutton, 1988), however that comes with the compromise that they are less flexible to change in the environment, and 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 instead of the state value function . This makes it possible to disassociate the value of different actions in a particular state , without the need to know how the states are interconnected.

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

The successful application of RL as a tool for neuroscience has been in part due to the discovery of phasic activity in the dopaminergic system analogous to the reward prediction error signals used in TD learning (Schultz et al., 1997). While these prediction error signals are an important basis for learning in model-free agents, the planning procedures associated with model-based methods do not explicitly necessitate a need for them. Despite this, multiple experiments designed to separate model-fee and model-based styles of learning have implicated dopaminergic prediction error as playing a causal role in model-based behaviour both in humans (Daw et al., 2011; Deserno et al., 2015; Wunderlich et al., 2012) and rodents (Sadacca et al., 2016; Sharpe et al., 2017). In light of this, recent years have seen a rise in application of the successor representation (SR) (Dayan, 1993). The SR functions somewhat as an intermediary between model-free and model-based methods (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:

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

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 decide actions.

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

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; I Momennejad et al., 2017) and rodents (de Cothi & Barry, 2020; Stachenfeld et al., 2017). Additionally, since the SR can be learnt using TD learning rules, it attributes a role for dopamine in forming sensory prediction errors (Gardner et al., 2018) - similar to what has been observed experimentally (Menegas et al., 2017; Sharpe et al., 2017; Takahashi et al., 2017).

The tasks often used to separate 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 conceptual nature of the underlying task space, translational research requires differing sensory implementations for humans (Daw et al., 2011) and rodents (Miller et al., 2017).

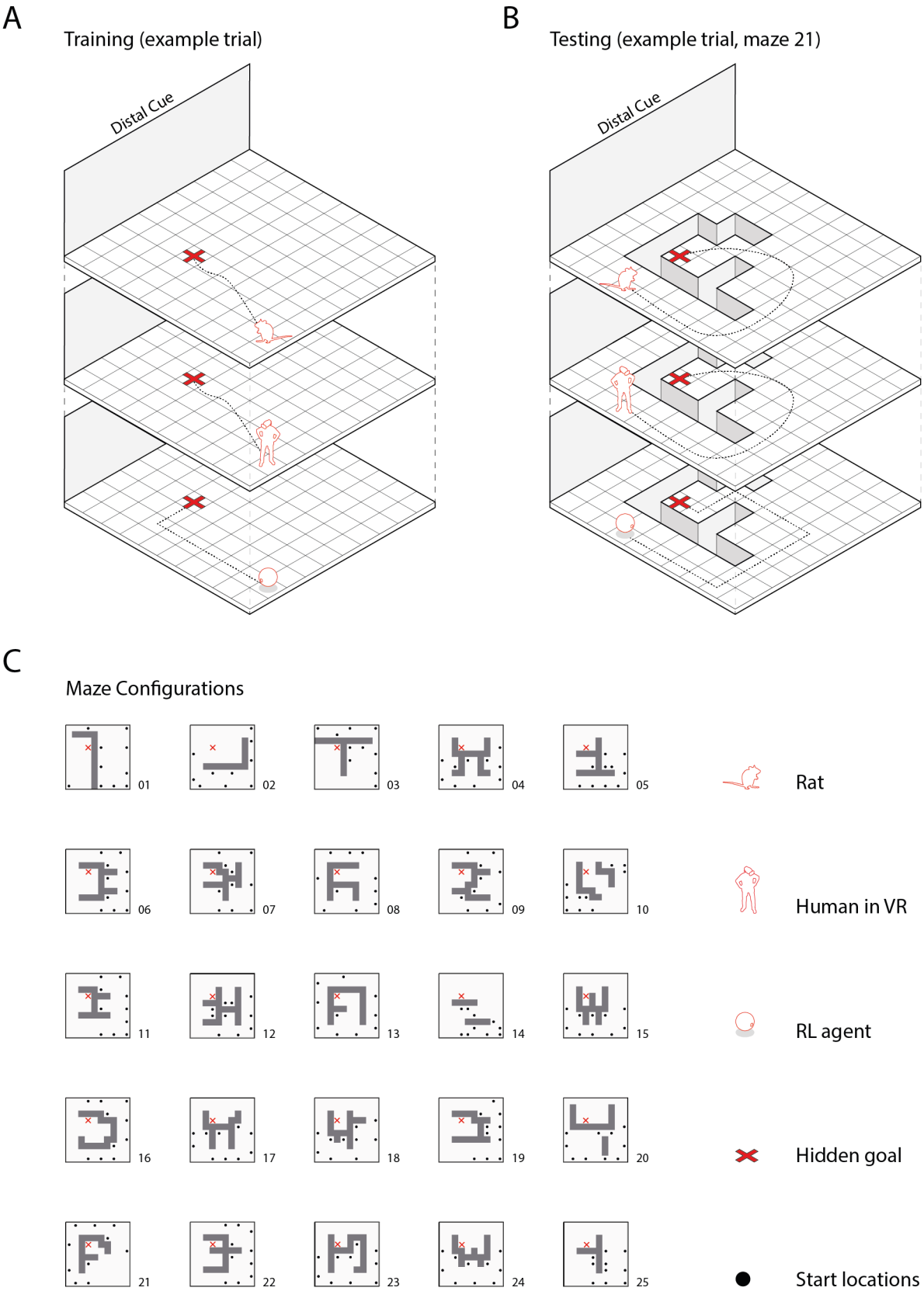
Here we exploited a range of analytic methods to investigate how rat and human spatial behaviours compare to model-free, model-based and SR reinforcement learners on a complex set of dynamic open-field mazes indicative of real-world navigation. We created a maze for rodents and an equivalent immersive VR for humans that could be used to directly compare their navigational behaviour and determine how it relates to model-free, model-based and SR RL agents.

Results

Navigation was tested in 2mx2m, 10x10 modular (20cmx20cm) flat open maze environment (60cm from the floor), with a fixed hidden goal (Fig. 1, Video S1). The modular maze design allowed sections of the 10x10 grid to be removed leaving impassable gaps in the floor to force detour taking or create shortcuts (when replaced). During training, all maze modules were present creating an open square arena. Rats, humans and RL agents were trained to navigate directly to the goal within a 45s time limit (Fig 1A). Crucially, during the testing phase of the experiment, maze modules were systematically removed in order 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 with the same 10 starting positions per maze (Fig 1C), enabling direct comparison of the resulting trajectories. These maze configurations were generated from a pilot testing with 9 rats and the sequence chosen maximized the differences in the layouts between trials. The starting positions on each maze gradually increased in the tortuosity (path distance / Euclidean distance) required of the shortest path to the goal in order to necessitate more complex trajectories whilst keeping the rodents motivated.

Navigation was facilitated by a large distal cue composed of a black curtain that spanned the majority of one edge of the maze (Fig 1A-B, Video S1). Upon reaching the goal module, rats and humans had to wait 5s to receive their reward. Human participants were rewarded with a financial bonus (supplementary methods) and rats received 0.1ml of chocolate milk delivered

upwards into a well in the floor of the maze goal module (See Extended Methods). 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. In order to better match the visual acuity and depth perception of rodents (Heffner & Heffner, 1992) to the human participants, a thick virtual fog lined the floor of the maze enabling them to only see adjacent maze modules. Modules were not visually made distinct to avoid humans counting module steps. Additionally, to match the instruction set between the human and the rodent experiments, participants were not given any specific information about the task before training and were only informed that reward was hidden in the environment and that their task was to maximise their financial return as quickly and efficiently as possible.



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

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

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 .

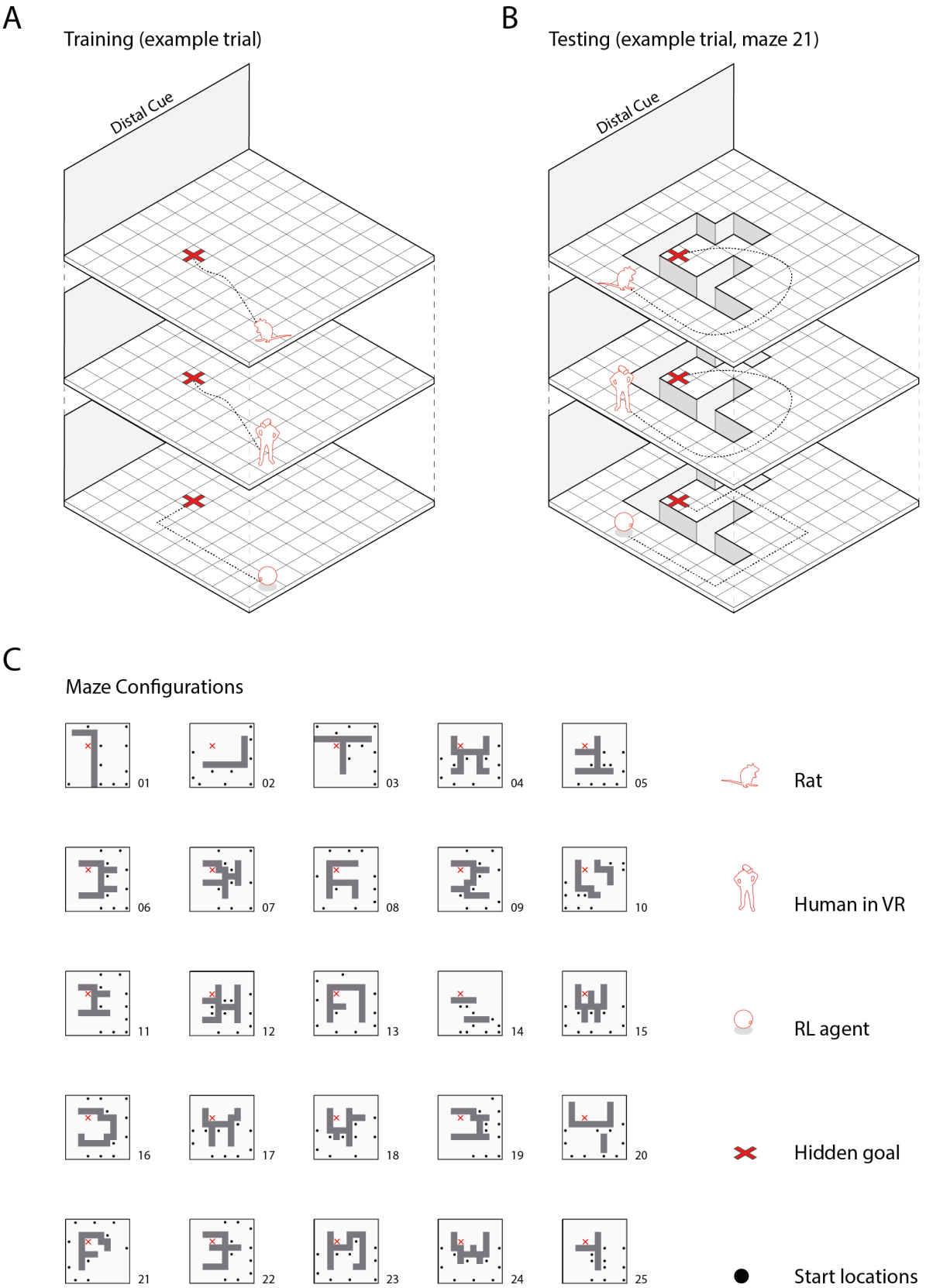


Figure 1: The Tartarus Maze. A) Humans, rats and reinforcement learning (RL) agents were trained to navigate to a hidden goal location (red 'x') within a 45s time limit. Maze diagram is illustrative and not to scale with the rat or human (see Supplemental Video 1 for rat trial performance). B) Parts of the maze floor were systematically removed or replaced every 10 trials in order to block the route to the goal and force detour taking or

the use of novel shortcuts. C) Humans, rats and RL agents were tested on the same sequence of 25 maze configurations with the same 10 starting positions in each maze.

Model-based agent

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

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

At decision time, the model-based agent uses planning procedure that follows the shortest route to the goal according to its model . This was implemented using an A* tree search algorithm (Hart et al., 1968).

Successor representation agent

The SR agent uses temporal-difference learning rules to update both successor matrix and state-rewards .

$$M(s,:) \leftarrow M(s,:) + \alpha\left[1_s + M(s',:) - M(s,:)\right] \quad (7)$$

$$R(s_{t+1}) \leftarrow R(s_{t+1}) + \alpha\left[r_t - R(s_{t+1})\right] \quad (8)$$

At decision time, the SR agent computes the value of prospective future states via equation 4, and under a greedy policy will choose the next state with the highest value. All internal representations used by the SR agent and both other RL agents were carried over onto the first trial of the subsequent maze.

The human and rat trajectories were discretised onto the underlying 10x10 modular grid (Fig 2A-B) in order to facilitate comparison between each other and the RL agents. We first sought to address how well the humans and rodents were able to complete the task. As expected, with longer exposure to a new maze configuration we saw 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 $t(17)=6.3$, $p=7.3\times10^{-6}$; rats $t(8)=4.0$, $p=4.2\times10^{-3}$). Humans were also generally better than the rats at finding the goal across the 25 mazes configurations (Fig 2D; humans vs rats: $t(25)=3.0$, $p=6.0\times10^{-3}$), although there were 3 mazes when this was not the case and the rats outperformed the humans (mazes 2, 10 and 19). We saw a strong correlation between the routes taken by the rats and human participants (occupancy correlation, humans vs rats: $\rho=0.67$), in particular towards the later trials on mazes (Fig 2E; occupancy correlations for first 5 trials vs last 5 trials: $t(8)=3.2$, $p=0.013$) when both were better at navigating to the goal. The routes used during these goal-reaching trajectories were also more efficient (Fig 2E; deviation from optimal path first 5 trials vs last 5 trials: human $t(17)=-5.0$, $p=1.1\times10^{-4}$; rats $t(8)=-4.0$, $p=4.0\times10^{-3}$)

with humans generally choosing more optimal routes than the rats (deviation from optimal path humans vs rats: $t(25)=-8.2$, $p=1.4\times10^{-8}$).

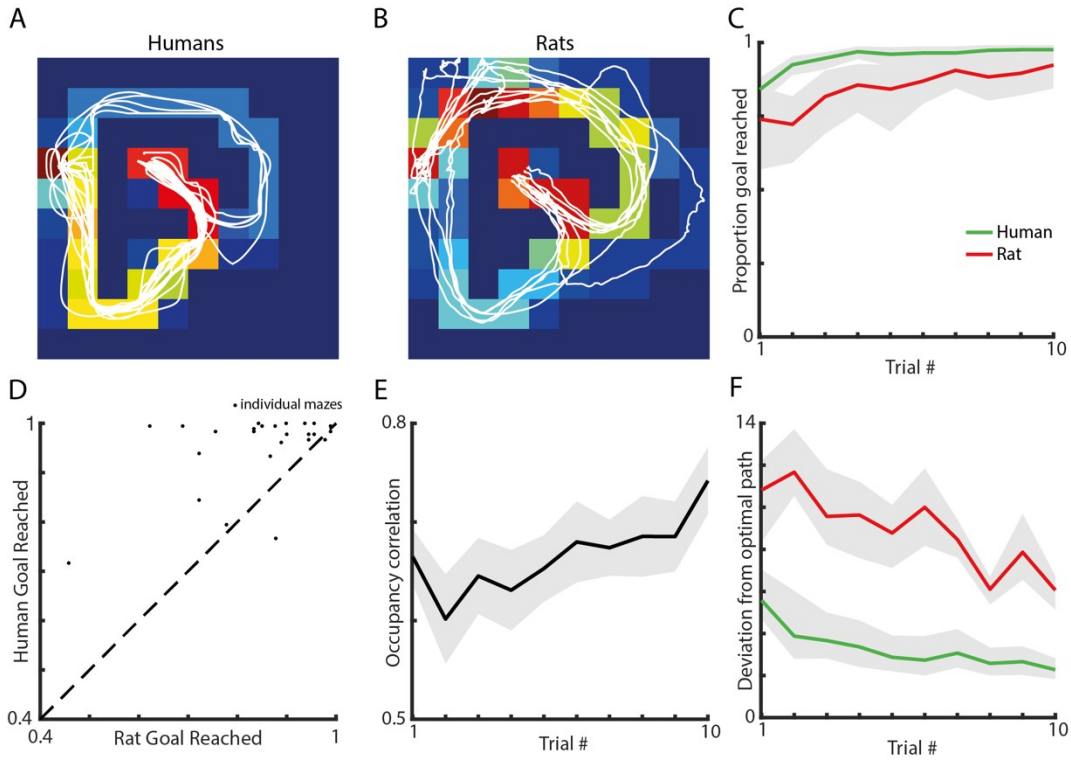


Figure 2: Humans and rats were able to successfully navigate to the hidden goal and did so using similar routes. Examples of the raw human (A) and rat (B) trajectories overlaying occupancy maps of the processed data used for analysis on a given trial (configuration 21, start 9). (C) The humans and rats are better at navigating to the hidden goal after more exposure to a maze configuration – although on the majority of mazes humans reach the goal more often than rats (D). (E) The correlation between the human and rat occupancy maps increases across exposure to a maze configuration implying they take increasingly similar routes - although the humans navigate to the goal with more efficient routes (F)

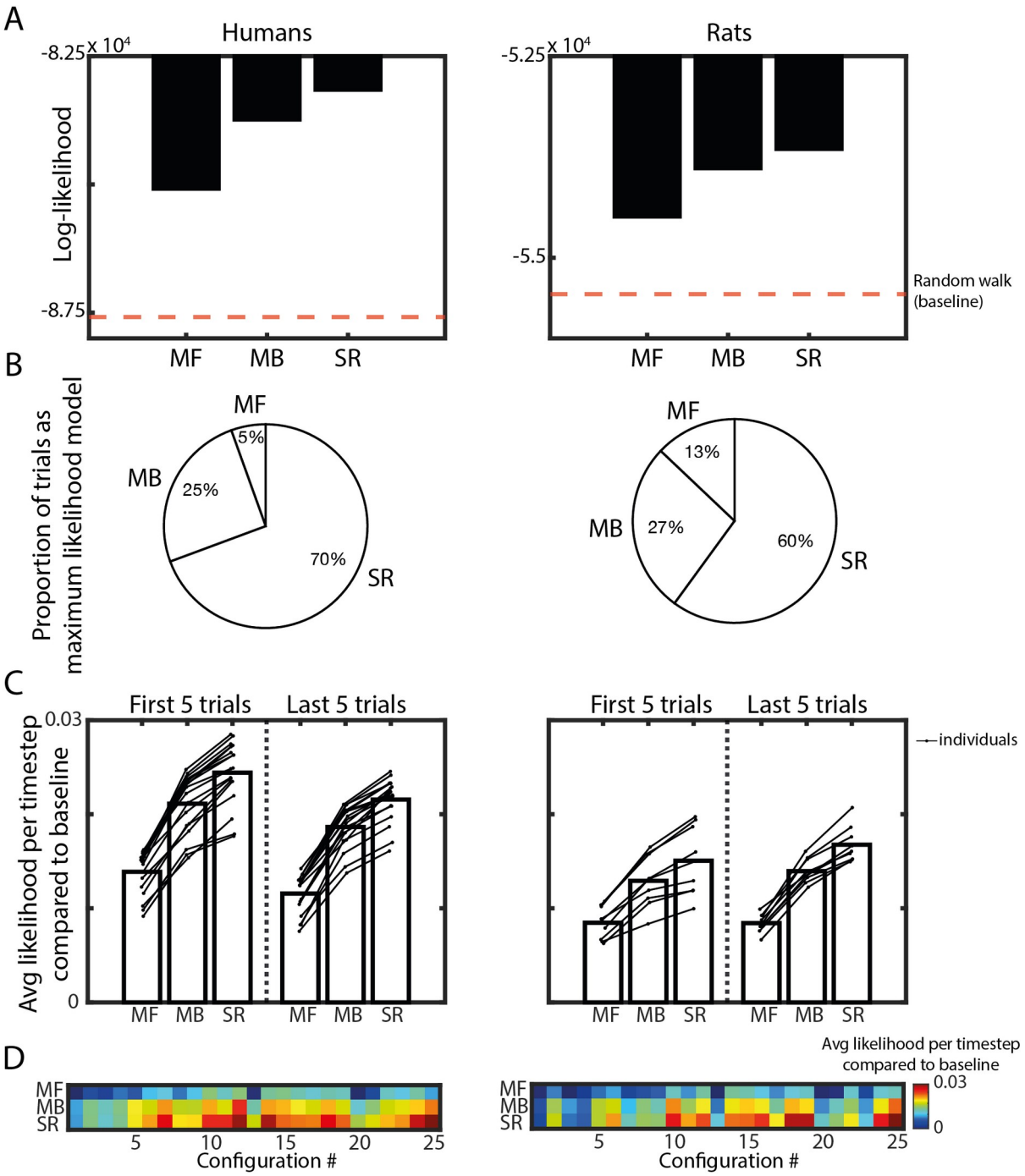
Likelihood analysis of actions reveals rats and humans are both most similar to an SR agent

We next investigated how the human and rat trajectories compared to the RL agents' trajectories. We used the human and rat behaviour 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 under 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).

Model Parameters				
	α		γ	
Model	Human	Rat	Human	Rat
MF	0.45	0.22	0.75	0.68
SR	0.93	0.81	0.79	0.79
MB	N/A	N/A	0.81	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; Bayes Factor (BF) analysis: SR vs MF for human data ln(BF) = 1911.1; SR vs MB for human data ln(BF) = 538.2; SR vs MF for rat data ln(BF) = 842.0; SR vs MB for rat data ln(BF) = 225.2), with the model-free agent consistently providing the worst fit (MF vs MB for human data ln(BF) = -1372.9; MF vs MB for rat data ln(BF) = -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 is robust throughout the time spent on a maze configuration (Fig 3C) and across individuals (SR vs MF for human data: $t(17) = 29.2$ $p=5.6\times10^{-16}$; SR vs MB for human data: $t(17) = 11.9$ $p=1.1\times10^{-9}$; SR vs MF for rat data: $t(8) = 13.0$ $p=1.1\times10^{-6}$; SR vs MB for rat data: $t(8) = 9.6$, $p=1.1\times10^{-5}$). 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 ($\rho=0.57$, $p=7.0\times10^{-8}$).



<u>Figure 3</u>: Maximum likelihood analysis of model-free (MF), model-based (MB) and successor representation (SR) reinforcement learning agents given the biological behaviour. (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 is trend is true across all individuals (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 ($\times$=0.57, p=7.0x10^{-8}).

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 actually transfer into differences in the resulting behaviour, we simulated the agent trajectories using the pooled maximum likelihood parameters (Table 1) and an epsilon-greedy policy with epsilon=0.1. This policy ensures that agents act greedily (i.e. choosing the action with maximum expected value) 90% of the time and choose actions randomly on the remaining 10% of time-steps in order to manage the exploration-exploitation dilemma. To accurately depict the distribution of trajectories generated by an agent under such a policy, we simulated each RL algorithm n=1000 times, with the maximum number of time-steps set to the mean unsuccessful trial length from the human and rat behaviour (human: T=42, rat: T=29).

The model-based algorithm was consistently the most successful agent at navigating to the goal (Fig 4A-B; Mann-Whitney test, human parameters: MB vs SR $z=29.6$ $p<0.001$, MB vs MF $z=38.7$ $p<0.001$; rat parameters: MB vs SR $z=14.5$ $p<0.001$, MB vs MF $z=38.7$ $p<0.001$), with the model-free agent consistently the worst at goal-reaching (human parameters: $z=-38.7$ $p<0.001$ MF vs SR; rat parameters: MF vs SR $z=-38.7$ $p<0.001$) – although the biological behaviour consistently outperformed all RL agents. As with the humans and rats, we see that the model-based and SR agents progressively improve at navigating to the goal throughout the trials on a maze configuration (Fig 4A-B; first 5 vs last 5 trials, human parameters: MB, $z=-18.1$, $p<0.001$; SR, $z=-30.1$, $p<0.001$; rat parameters: MB $z=-14.2$ $p<0.001$, SR: $z=-11.0$ $p<0.001$). Meanwhile, the model-free agents got progressively worse at reaching the goal (first 5 vs last 5 trials: human parameters, $z=38.6$, $p<0.001$; rat parameters, $z=38.8$, $p<0.001$) – indicative of the increasingly more tortuous starting locations. 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 reveals a significant positive correlation between the human and rat difficulty rankings with the SR agent (Fig 4E-F; human-SR rank correlation: $\rho=0.52$, $p=8.3\times10^{-3}$; rat-SR rank correlation: $\rho=0.40$, $p=0.049$). No significant correlations were detected between the human and rat difficulty rankings with either of the other agents (human-MF rank correlation: $\rho=0.15$, $p=0.48$; human-MB rank correlation: $\rho=-0.18$, $p=0.39$, rat-MF rank correlation: $\rho=0.37$, $p=0.066$; rat-MB rank correlation: $\rho=0.085$, $p=0.69$).

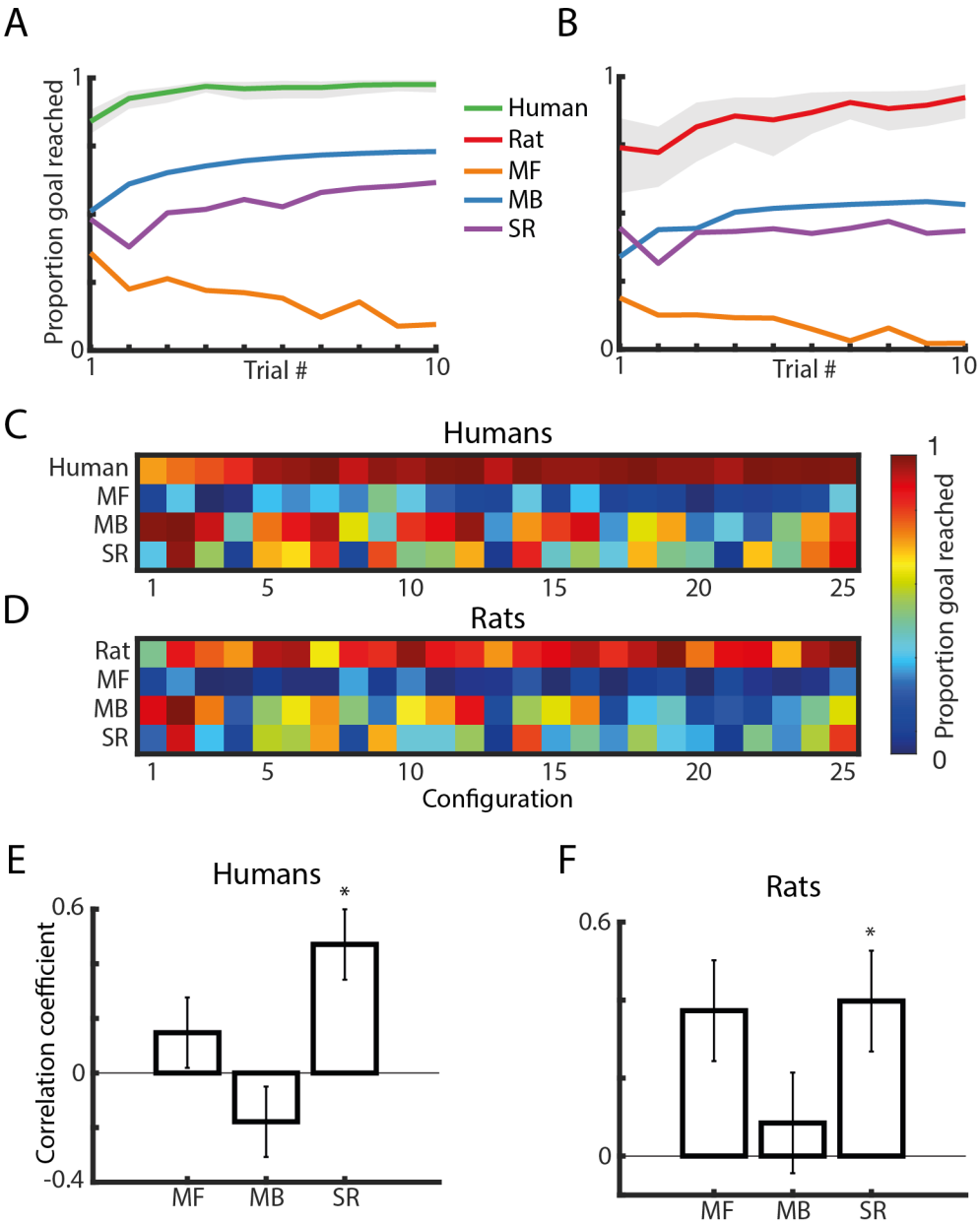


Figure 4: Comparison of human and rat behaviour with maximum likelihood reinforcement learning agents. Goal reaching for agents (n=1000) using the maximum likelihood parameters fit to the human (A) and rat (B) trajectories. (C-D) Goal reaching varied across maze configurations. Using this to rank maze configurations difficulty reveals a significant positive correlation between the human (E) and rat (F) difficulty rankings and the successor representation agent's (humans, $\times$=0.52, p=8.3x10^{-3}; rats, $\times$=0.40, p=0.049), but not with the model-free (humans, $\times$=0.15, p=0.48; rats, $\times$=0.37, p=0.066) or model-based (humans, $\times$=-0.18, p=0.39; rats, $\times$=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.

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; angular diffusivity and tortuosity. (Fig 4A). Linear and angular diffusivity describe the directness and direction of the route, whilst tortuosity describes the general trajectory complexity.

Visualising the quantified trajectories for a specific trial reveals distinct clustering of behaviour (Fig 5E). We then used the Mahalanobis distance to calculate the dissimilarity between the biological and agent trajectories. In particular, 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: $t(17)=-26.3$, $p=3.2\times10^{-15}$; rats to human cluster vs SR cluster: $t(8)=-6.0$, $p=3.3\times10^{-4}$). Of the RL agents, the SR was the most similar to the rat (SR vs MF: $t(8)=-27.5$, $p=3.3\times10^{-9}$; SR vs MB: $t(8)=-34.6$, $p=5.4\times10^{-10}$) and human trajectories (SR vs MF: $t(17)=-97.7$, $p=8.0\times10^{-25}$; SR vs MB: $t(17)=-2.4$, $p=0.03$), with the model-free agent consistently being least similar (humans to MF cluster vs MB cluster: $t(17)=37.6$, $p=8.4\times10^{-18}$; rats to MF cluster vs MB cluster: $t(8)=20.7$, $p=3.1\times10^{-8}$).

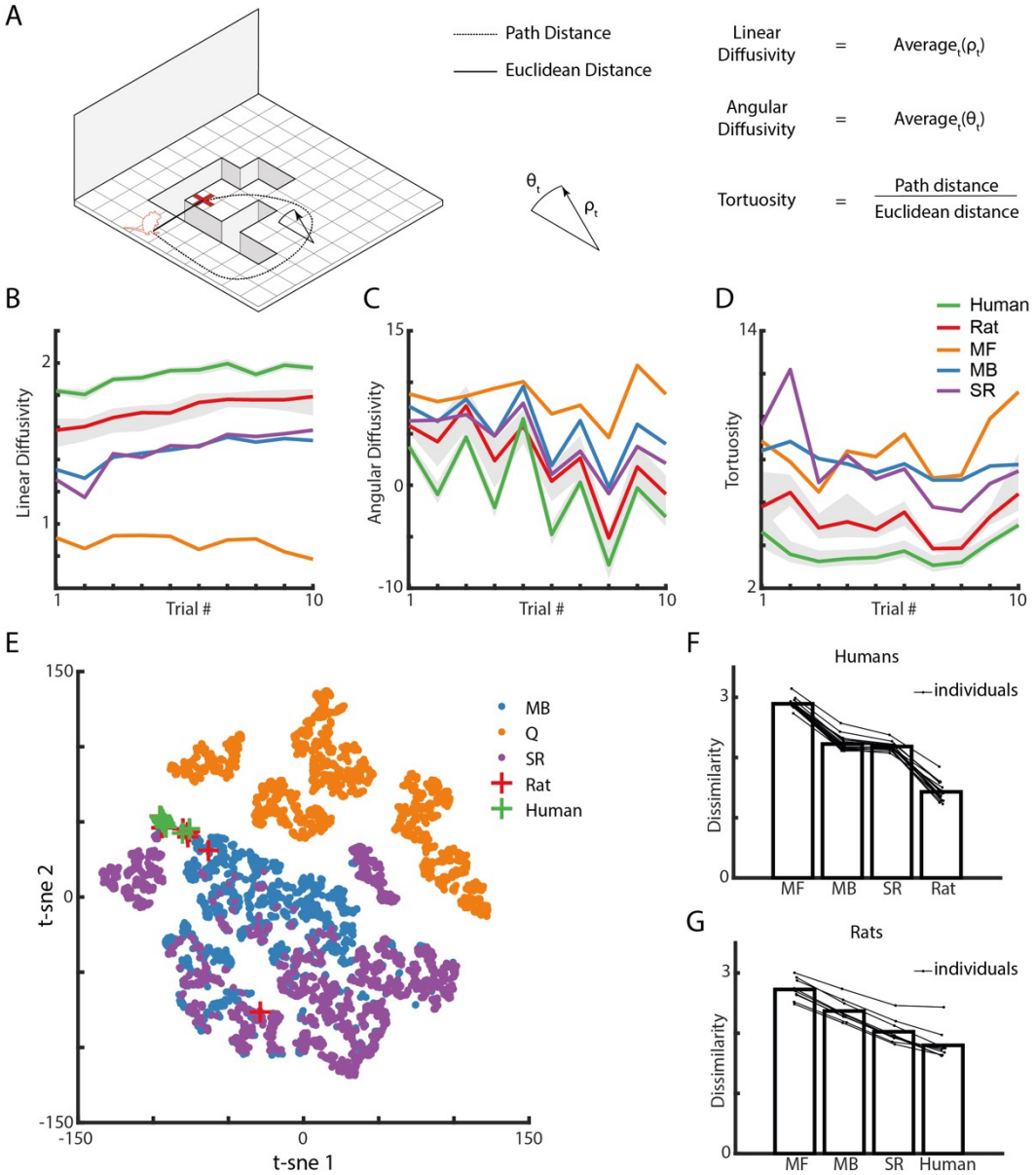


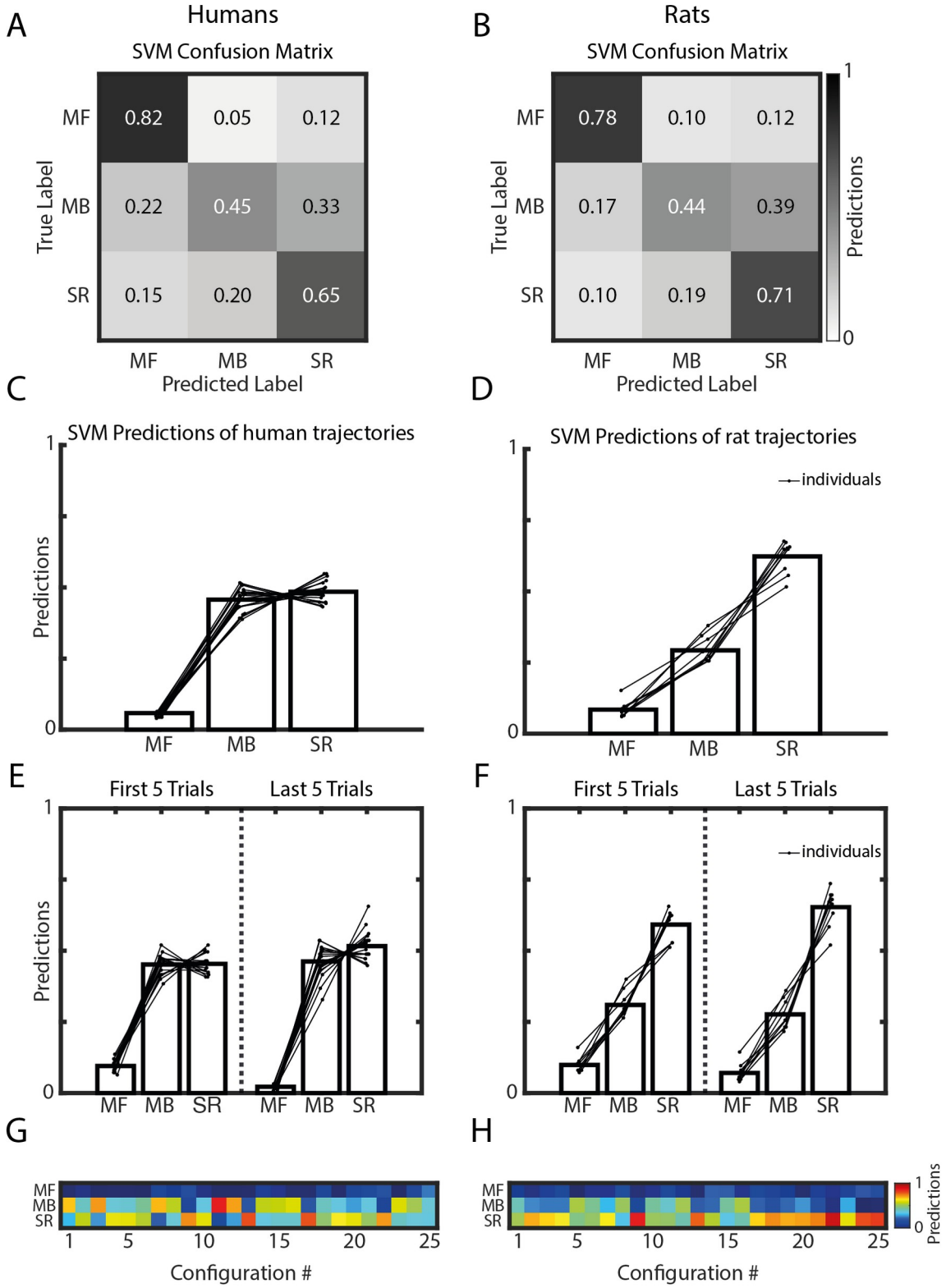
Figure 5: Similarity analysis reveals rat trajectories are most similar to a successor representation agent, whilst human trajectories are similar to both successor representation and model-based agents. (A) The shape of each trajectory was quantified using three metrics: the average diffusivity (B) and average

angular velocity (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. 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 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 (Fig 6A-B). When the decoder was given data from the biological behaviour, the SVM classifier most frequently predicted those trajectories to be an SR agent (Fig6CD). This was most pronounced for the SVM classification of the rodent data (SR vs MB: $t(8)=9.8$, $p=1.0\times10^{-5}$, SR vs MF: $t(8)=11.8$, $p=2.4\times10^{-6}$), with no discernible difference in SR and model-based agent predictions until the latter half of trials for the human participants (Fig6EF; first 5 trials $t(17)=0.14$, $p=0.89$; last 5 trials $t(17)=2.1$, $p=0.048$). Viewing how SVM predictions change from maze-to-maze again reveals noticeable variation (Fig 6GH), with configurations that have little overlap in policy with the preceding maze often leading to increased SVM predictions of the SR agent.

In summary, using three approaches to compare RL agents to rats and humans our results show that both species match more closely the SR RL agents performance patterns than MF or MB agents.



<u>Figure 6</u>: Support vector machine (SVM) classifies rats trajectories as SR agents and humans as SR or MB agents. The SVM classifier is 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. The SVM classifier most often classifies the human trajectories (C) as MB and SR agents (SR vs MB: t(17)=1.51, p=0.15), whilst the rat behaviour (D) is most commonly predicted as an SR agent (SR vs MB: t(8)=9.8, p=1.0x10-5). Later trials on a configuration are more likely to be classified as an SR agent for both the human (E; SR vs MB prediction on last 5 trials of maze,

t(17)=2.1, p=0.048) and rat data (F; SR vs MB prediction on last 5 trials of maze, t(8)=11.1, p=3.7×10^{-6}). The level of these predictions varied from maze-to-maze (G-H) with a strong correlation between the humans and rats (rank correlation: rho = 0.71, p = 1.49e-12).

Discussion

To understand the underlying processes that support flexible navigation in rats and humans we compared their performance on a novel navigation task (Tartarus Maze) against three classic instantiations of RL agents in simulated versions of the same task. By using a combination of likelihood analysis, trajectory similarity analysis and a support vector machine classifier we find that both species produce similar trajectory patterns, which are most similar to successor representation RL agents, with humans also showing some similarity to model-based RL agents in their trajectory patterns. To our knowledge, this is the first translational investigation into the spatial navigation strategies of humans and rats that directly comparing them on the same task to reinforcement learning agents.

Previous experiments comparing human behaviour to RL agents have focused on a dichotomy of model-free and model-based behaviours in small conceptual state spaces with 2-step transitions and found evidence for model-based planning in humans (Daw et al., 2011; Vikbladh et al., 2019; Wunderlich et al., 2012). By using a larger state space with potential for recursive transitions we extend this into a more complex and naturalistic framework. Our finding build upon recent work showing that the actions of human participants are best explained by a combination of SR and model-based behaviours (I 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 provide a function for hippocampal replay (Mattar & Daw, 2018; Ida Momennejad, 2020; Ida Momennejad et al., 2018; Ólafsdóttir et al., 2017; Pfeiffer & Foster, 2013). 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; Stachenfeld et al., 2017) by showing it also 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.

Prior studies examining rodent navigation behaviour have generally used either tracks or open field environments (Poulter et al., 2018). While open field environments place demands on self-localisation and vector-based navigation, they do not typically allow examination of adaptive behaviour. Mazes composed of tracks can nicely examine blocked paths (Grieves & Dudchenko, 2013; E C Tolman & Honzik, 1930), but put less demands on self-localisation and vector-based navigation. By contrast our 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. By having the entire path to the goal available from the outset we were able to examine ballistic executions of chosen paths. 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 have tended to involve VR environment more complicated than the current example (Javadi, Patai, Marin-Garcia, Margois, et al., 2019; Javadi, Patai, Marin-Garcia, Margolis, et al., 2019). Thus, our results show the potential for studing navigation with a task that can be matched across rats and humans.

Humans performed more accurately than rats in our task. There may be multiple reasons for this, including the motivational demands (money vs food), the differences in the exact presentation of the maze (scale, view, speed of movement, etc) and the key fact that humans likely understand the task set for them by the experimenters. For these reasons it is difficult to interpret the differences in the behaviour of the two species. Nonetheless it is notable that rats outperformed humans when navigating in configuration 2 where the goal is not near a wall. It is likely that rats were using olfactory cues to aid goal localisation near the goal, which may have aided navigation, but the speed of travel across the routes suggest predominantly visual cue guided navigation.

There was a noticeable difference in successful goal-reaching between the biological behaviour and RL agents. It has been shown that the learning efficiency of RL agents can be improved using offline replay of randomly sampled past experiences (Russek et al., 2017; Sutton, 1990). 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) – although agent replay was not implemented in this study in order to keep the value representations and consequently the likelihoods of agents deterministic. An alternative way to improve the goal-reaching of agents could be through improving their exploration policy. The agents simulated here implemented 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.

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 the 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 distances 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). Whilst averaged we averaged the diffusivity measures across the trajectory, they could also be calculated online to identify different modes during actively tracked navigation.

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.

References

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

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, Issue 4, pp. 603–616). Cell Press. https://doi.org/10.1016/j.neuron.2020.06.014

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

Clemens, L. E., Jansson, E. K. H., Portal, E., Riess, O., & Nguyen, H. P. (2014). A behavioral comparison of the common laboratory rat strains Lister Hooded, Lewis, Fischer 344 and Wistar in an automated homecage system. *Genes, Brain and Behavior*, *13*(3), 305–321. https://doi.org/10.1111/gbb.12093

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, Issue 4, pp. 429–453). Springer. https://doi.org/10.3758/CABN.8.4.429

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

Deserno, L., Huys, Q. J. M., Boehme, R., Buchert, R., Heinze, H. J., Grace, A. A., Dolan, R. J., Heinz, A., & Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(5), 1595–1600. https://doi.org/10.1073/pnas.1417219112

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

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. In *Nature Neuroscience* (Vol. 20, Issue 11, pp. 1504–1513). Nature Publishing Group. 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

Gardner, M. P. H., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1891). https://doi.org/10.1098/rspb.2018.1645

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

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

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

Harris, M. A., & Wolbers, T. (2012). Ageing effects on path integration and landmark navigation. *Hippocampus*, *22*(8), 1770–1780. https://doi.org/10.1002/hipo.22011

Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, *4*(2), 100–107. https://doi.org/10.1109/TSSC.1968.300136

Heffner, R. S., & Heffner, H. E. (1992). Visual factors in sound localization in mammals. *The Journal of Comparative Neurology*, *317*(3), 219–232. https://doi.org/10.1002/cne.903170302

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*, 14652. 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*. https://doi.org/10.1162/jocn_a_01414

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

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

Menegas, W., Babayan, B. M., Uchida, N., & Watabe-Uchida, M. (2017). Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *ELife*, *6*. https://doi.org/10.7554/eLife.21886

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, 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

Momennejad, Ida. (2020). Learning Structures: Predictive Representations, Replay, and Generalization. In *Current Opinion in Behavioral Sciences* (Vol. 32, pp. 155–166). Elsevier Ltd. https://doi.org/10.1016/j.cobeha.2020.02.017

Momennejad, Ida, 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

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map.* Oxford: Clarendon Press. https://doi.org/10.1097/00005053-198003000-00018

Ólafsdóttir, H. F., Carpenter, F., & Barry, C. (2017). Task Demands Predict a Dynamic Switch in the Content of Awake Hippocampal Replay. *Neuron*, *96*(4), 925-935.e6. https://doi.org/10.1016/j.neuron.2017.09.035

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

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

Sadacca, B. F., Jones, J. L., & Schoenbaum, G. (2016). Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *ELife*, *5*(MARCH2016). https://doi.org/10.7554/eLife.13665

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., Niv, Y., & Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, *20*(5), 735–742. https://doi.org/10.1038/nn.4538

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

Spiers, H. J. (2008). Keeping the goal in mind: Prefrontal contributions to spatial navigation. *Neuropsychologia*, *46*(7), 2106–2108. https://doi.org/10.1016/j.neuropsychologia.2008.01.028

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

Spiers, H. J., Hayman, R. M. A., Jovalekic, A., Marozzi, E., & Jeffery, K. J. (2015). Place field repetition and purely local remapping in a multicompartment environment. *Cerebral Cortex*, *25*(1), 10–25. https://doi.org/10.1093/cercor/bht198

Spiers, H. J., & Maguire, E. A. (2006). Thoughts, behaviour, and brain dynamics during navigation in the real world. *NeuroImage*, *31*(4), 1826–1840. https://doi.org/10.1016/j.neuroimage.2006.01.037

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. (1990). Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. In *Machine Learning Proceedings 1990* (pp. 216–224). Elsevier. https://doi.org/10.1016/b978-1-55860-141-3.50030-4

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

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M., & Schoenbaum, G. (2017). Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. *Neuron*, *95*(6), 1395-1405.e3. https://doi.org/10.1016/j.neuron.2017.08.025

Tolman, E C, & Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology*, *4*, 257–275.

Tolman, Edward C. (1948). Cognitive maps in rats and men. In *Psychological Review* (Vol. 55, Issue 4, pp. 189–208). American Psychological Association. https://doi.org/10.1037/h0061626

Uylings, H. B. M., Groenewegen, H. J., & Kolb, B. (2003). Do rats have a prefrontal cortex? In *Behavioural Brain Research* (Vol. 146, Issues 1–2, pp. 3–17). Elsevier. https://doi.org/10.1016/j.bbr.2003.09.028

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

Wood, R. A., Bauza, M., Krupic, J., Burton, S., Delekate, A., Chan, D., & O'keefe, J. (2018). The honeycomb maze provides a novel test to study hippocampal-dependent spatial navigation. *Nature Publishing Group*, *554*. https://doi.org/10.1038/nature25433

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., & Håberg, 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

Gahnstrom & Spiers (2020) *under review*

Supplementary Methods

In all versions of the experiment, the environment 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

Reinforcement learner simulations

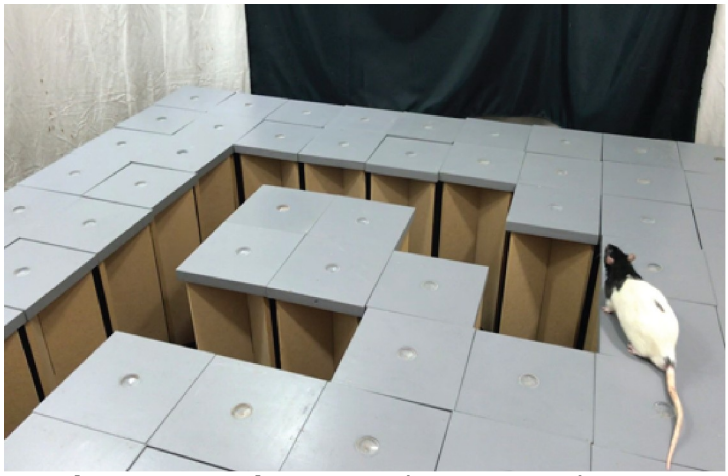
The reinforcement learning agents were implemented in a 10x10 grid world. At the beginning of the experiment, all agents were endowed with the optimal policy on the open maze to simulate the training phase undertaken by rats and humans. They were then run consecutively on the 25 maze configuration, carrying 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 $\varepsilon = 0.1$. 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.

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 naïve, 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 via two pieces of wood supports fit together through their long slits (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, 9mm depth) at its centre, which could be attached to polymeric tubing system installed beneath the maze. This tubing allowed the experimenter to reward the rat at the goal module by soundlessly filling the well with chocolate milk (0.1 ml). 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. The goal module was always in the same position with respect to this cue.



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.

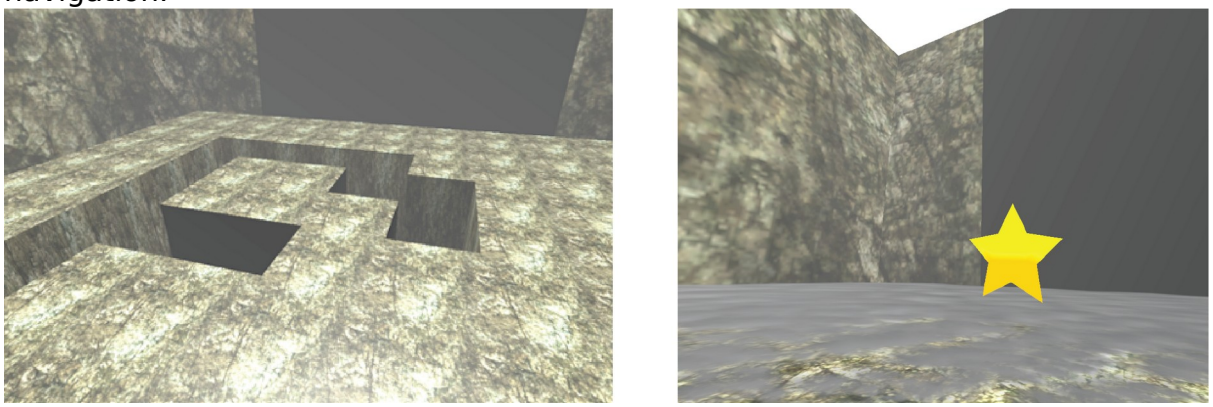
An initial familiarisation phase lasted for three days. During the first day, the rats received a small amount (0.1ml per rat) of chocolate milk in the home cage to decrease neophobia on the maze. For the subsequent two days, each rat underwent two 15 minute maze familiarisation sessions, in which the 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. The experimenter was present beside the maze inside the curtained area throughout the session, and between sessions the black sheet was rotated 90° counter-clockwise.

After the familiarisation phase, rats began to be trained to navigate to the goal location. In each training trial the rat had 45s to find the goal module, during which the experimenter stood motionless next to the maze. Training consisted of three stages. The first stage lasted one day and consisted of two 15 minute sessions, during which the goal module's well was filled with 0.1ml of chocolate milk and the rats were initially placed on the modules adjacent to the goal. For each subsequent training trial, the rat's starting position would be shifted one module anticlockwise. 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. In this first stage, the rats were always placed facing the goal. The second training stage followed the same procedure was as stage one, but 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 edges of the maze. The third and final training stage was also similar to stages one and two, except the number of daily trials could be increased up to 25. Furthermore, the rat's starting position and orientation was randomised and 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. 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

Following the training phase, rats were run on the 25 maze configurations 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. At the beginning of each day, rats were given a brief reminder session that consisted of 5 trials from phase 3 of the training phase, and then the rats would usually go on to complete 3 configurations per day.

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 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) 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 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.