Individual differences in model-based planning are linked to the ability to infer latent structure

Milena Rmus[1], Harrison Ritz[2], Lindsay E Hunter[3], Aaron M Bornstein[4,5], Amitai Shenhav[2,6]

[1] *Department of Psychology, University of California, Berkeley*

[2] *Department of Cognitive, Linguistic, and Psychological Sciences, Brown University*

[3] *Department of Psychology, Princeton University*

[4] *Department of Cognitive Sciences, University of California, Irvine*

[5] *Center for the Neurobiology of Learning and Memory, University of California, Irvine*

[6] *Carney Institute for Brain Science, Brown University*

Abstract

To behave adaptively, people must choose actions that maximize their expected future rewards. Engaging in such *goal-directed* decision-making in turn requires the capacity to (1) develop an internal model of one's environment (i.e., representing the relationship between current and future states; *structure inference*), and (2) navigate this cognitive model to determine the action(s) that will lead to the most rewarding future state (*model-based planning*). While previous work has identified putative mechanisms underlying these two processes, it has yet to test the prediction that one's ability to infer structure should constrain their ability to engage in model-based planning. Here we test this prediction using a novel task we developed to specifically isolate individual differences in structure inference ability. Participants (N=77) viewed a series of object pairs. Unbeknownst to them, each pair was drawn at random from adjacent nodes in an underlying graph. They then performed two tasks that measured the extent to which participants were able to infer the graph structure from these disjointed pairs: (1) judging the relative distances of sets of three nodes, (2) constructing the graph. We identified a single underlying factor that captured variability in performance across these tasks, and showed that this variability in this measure of structure inference ability was selectively associated with the extent to which participants exhibited model-based planning in the two-step task (Daw et al., 2011), a well-characterized assay of such behavior. Our work validates a new method for isolating one's capacity for structure inference, and confirms that individuals who are more limited in this capacity are less likely to engage in model-based planning. These findings bridge separate areas of research that examine goal-directed planning and its component processes. They further provide a path towards better understanding deficits in these component processes, and how they constrain one's ability to achieve long-term goals.

Humans have a remarkable ability to construct complex, goal-directed plans. We can plan the steps needed to complete a multi-part task; plan the words we will use to communicate a new idea; plan a route through an unfamiliar city; or plan an event several months or even years away. Achieving these goals relies on two component processes. First, we need to infer the *structure* of a given environment, including how to get between different states in that environment (e.g., different locations in space or different steps in a task sequence). Second, we need to generate and implement a *plan*: a sequence of actions that leverages this internal model of the environment in the service of a particular goal. These two processes are jointly necessary for successful goal-directed behavior, but have not yet been separately measured. As a result, it is not yet known whether the ability to construct such internal models based on one's experience with an environment (*structure inference*) entails the ability to use those models to achieve the best outcome (*model-based planning*). Here, we introduce and validate a task that separately measures structure inference ability, and test whether individual differences in this ability predict the use of model-based planning.

One body of work has examined how people develop internal models of their environment based on their experience with individual states in that environment and the transitions between them (Fermin et al. 2010; Behrens et al. 2018). Foundational research in the area demonstrated that animals construct *cognitive maps* as they navigate their spatial environment (Tolman, 1948; O'Keefe and Nadel, 1978), and that neural representations of these maps (decoded from regions of hippocampus) not only reflect the animal's location in that space but also (1) their recent locations and (2) the future projections of locations they intend to visit (Johnson & Redish, 2007). Recent work has shown that cognitive maps can also be extrapolated from *abstract* learned associations. For instance, Schapiro and colleagues (2013) built a virtual graph-like structure, with each node represented by an individual abstract stimulus. In their experiment, participants traversed this graph sequentially, one node at a time. Despite never seeing the underlying graph, participants were able to recover the graph, based on their experience of the likelihood of moving from one node to another. In addition, much like the representation of spatial maps, the graph representation itself could also be decoded from their brain activity. Similar forms of construction and navigation have been demonstrated over episodic and semantic memory representations, including connections between words (Jurafsky, 1996), concepts (Collins & Quillian 1969), events (Collin et al. 2015), and people (Parkinson et al., 2017; Tamir & Thornton, 2018; FeldmanHall & Shenhav, 2019).

A second body of work has examined the process by which people navigate these internal models in order to determine the course of action that will maximize their future rewards (Sutton and Barto 1998; Daw et al. 2005; Daw et al. 2011). Early work demonstrated that these goal-directed forms of decision-making trade off against habitual behavior, enabling an animal to adapt to rapid changes in their environment (Balleine & Dickinson 1998; Balleine & O'Doherty, 2009). Distinct neural circuits were shown to be causally involved in making an animal more or less goal-directed, influencing, for instance, the extent to which they maintained a previous course of action after it was no longer rewarded or no longer necessary to obtain a given reward (Yin et al. 2004; Yin et al. 2005; Yin et al. 2006). More recently, it has been shown that goal-directed decision-making can be formalized as *model-based reinforcement learning* (RL) (Daw et al., 2005; Sutton and Barto,1998; Dolan & Dayan, 2013). Model-based RL represents a form of RL that stores a model of how different states in the environment are connected to one another (e.g., the likelihood of transitioning from one state to another), and the rewards that an agent can expect upon reaching a given state. By contrast, *model-free* forms of RL only store the value of previous actions taken in a given state, and therefore are less sensitive to changes in the structure of one's environment (e.g., if a certain state is no longer rewarded or if two states are no longer connected, requiring a detour).

These two types of RL - model-based and model-free - are commonly dissociated with a task developed by Daw and colleagues (2011), in which participants must choose between a pair of initial states (States A and B), and based on their choice then transition to one of two new states (States C and D). At the new state, participants choose between two options (C1 vs C2 in State C, or D1 vs. D2 in State D) and either receive reward or do not. Critically, the probability of transitioning between States A/B and States C/D is not deterministic. Instead, each of the initial states is more likely to lead to one of the states at the next stage (e.g., State A to State C) but sometimes leads to the other state (e.g., State D). Upon receiving reward based on their second-stage choice (e.g., C1), a model-free agent will choose the initial state that previously landed them in State C (either State A or State B), whereas a model-based agent will choose the state that is *on average most likely* to land them in State C (e.g., State A). Patterns of first-stage decisions on this "two-step task" can therefore reveal the extent to which a participant engages in *model-free* decision-making - choosing actions based only on whether they were recently rewarded - or *model-based* planning - choosing actions based on a consideration of

both the recent rewards and the likelihood of reaching those rewards given the transition structure of the task environment (Daw et al., 2011; Decker et al., 2016). Using this task, researchers have shown that individual differences in one's tendency to engage in model-based decision-making have been linked to variability in working memory capacity (Otto et al. 2013), cognitive control (Daw, Niv & Dayan 2005; Otto et al. 2015), temporal discounting (Shenhav, Rand & Greene 2016; Hunter, Bornstein, Hartley 2019) and psychiatric symptoms associated with compulsive behavior and social isolation (Gillan et al., 2016).

Goal-directed planning thus depends critically on both our ability to (1) learn the structure of one's environment *and* (2) our ability to leverage the representation of this structure in pursuit of rewards. Recently, a consensus has developed that these capacities share overlapping computational and neural substrates (Collin et al. 2015; Shohamy & Turk-Browne 2013; Behrens et al. 2018; Vikbladh et al. 2019). However, while the mechanisms that support learning, navigating, and deploying an internal model have separately been well-characterized, the relationships between these domains remain poorly understood. In particular, work on model-based planning uses tasks in which the associative structure is made explicit, which de-emphasizes structure learning, leaving behavioral measures of goal-directed behavior to index how these representations are used. As a result, little is known about how one's ability to infer the structure of an environment relates to, and perhaps constrains, their ability to leverage such a representation when engaging in goal-directed planning. Here, we test this relationship directly, by examining whether a person's ability to construct and navigate a cognitive map predicts the degree to which they engage in model-based decision-making. We developed a novel set of tasks to measure participants' ability to infer the structure of an abstract (non-spatial) graph, based on disjoint experiences with pairs of adjacent nodes throughout that graph. We found that participants who exhibit better structure inference abilities were also more likely to engage in model-based planning in the two-step task. This work firmly connects these two cognitive functions, while also validating a novel approach to measuring individual differences in the ability to infer and navigate latent structures in one's environment.

Results

Participants (N=77) performed a novel structure inference and judgement task with three main phases (Figure 1). In Phase 1, participants were given the opportunity to implicitly learn a graph-like structure through experiences with pairs of nodes in that graph. On each of 704 trials,

they viewed a pair of objects (e.g., a bowl and a clipboard) and asked to report whether one of the objects was rotated relative to a canonical orientation (Figure 1B). Though they were never informed of this, each of these object pairs reflected a randomly drawn pair of adjacent nodes from an underlying graph, for which each node was represented by a single object (e.g., a clipboard) (Figure 1A). In Phase 2, participants made a series of judgments about the relative distances of three randomly selected nodes in the graph. On each trial, they were asked to evaluate which of two objects they thought was "closer" to a third reference object (Figure 1C), requiring them to make implicit inferences about the latent structure of the graph. In Phase 3, participants were asked to freely arrange the 12 objects and their connections in order to reconstruct their best estimate of the underlying graph (Figure 1D).

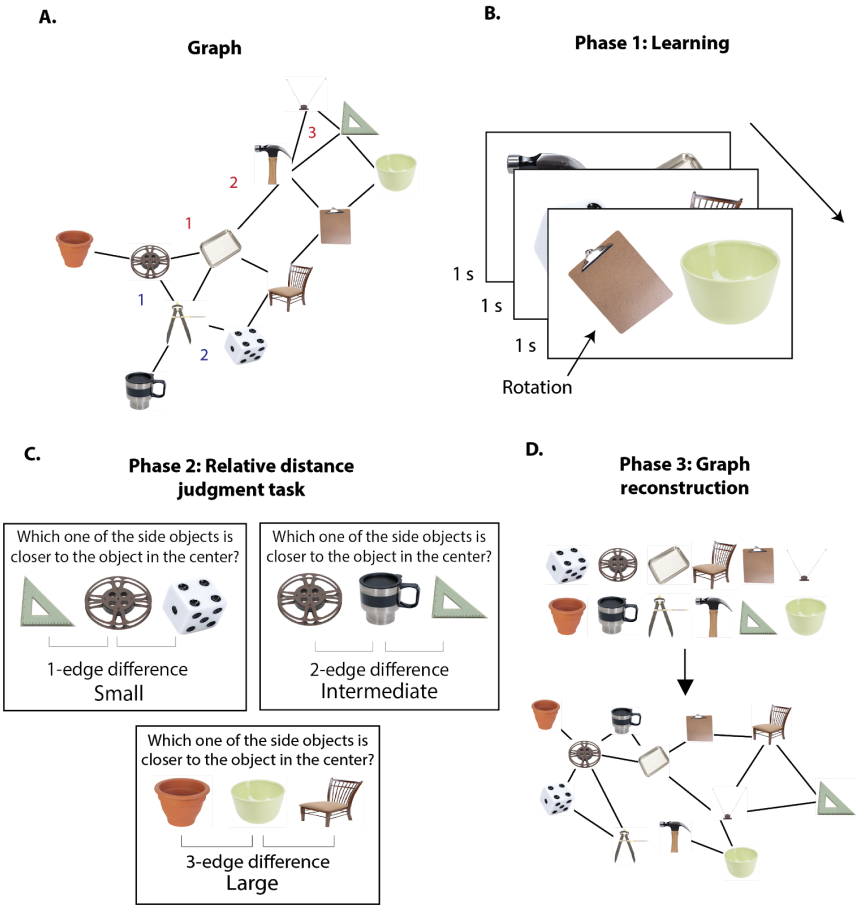

Figure 1. Schematic of the graph task. **A)** Underlying graph. Node labels (the object images) were fully randomized across all subjects. **B)** Learning phase: Participants observed pairs of randomly sampled adjacent nodes. **C)** Relative distance judgment task trial: participants were asked to identify the more proximal node. **D)** Graph reconstruction: participants freely arranged objects and connections between them, based on the learned object associations.

Evidence of structure inference

In Phase 1, participants performed very well on the rotation detection task (mean accuracy= 89%, SD= 5%; d'= 2.65, SD= .8; false alarm rate= .5%), indicating that they sustained attention throughout the learning phase. We then examined two sets of measures to assess whether participants were able to infer the underlying graph structure based on this learning phase.

Despite only experiences with disjoint node pairs from and underlying graph that they were not made explicitly aware of, participants were able to discern the distances between nodes they had never seen paired together, as indicated by above-chance performance on the relative distance judgment task in Phase 2 (mean accuracy= 67%, SD= 13%; t(76)=11.49, p= 2.5e-18). Importantly, these distance judgements were also sensitive to the overall difficulty of the distance judgement: participants were both faster (Figure 2; β =-.03 (.01), t= -2.30, p = .02) and more accurate (Figure 2; β=.02 (.008), t= 3.50, p = .0004) depending on how much closer the reference was to the target, relative to the foil. Response times reflected the relative time it might have taken participants to search for one node relative to the other (from the reference node). If so, the RTs should not only scale with the relative distance between the nodes, but also with the *total* distance from the reference to the two other nodes (e.g., the depth of search). Consistent with this prediction, participants' response times were longer when the total distance of both options to the target was greater (β=.13, t=7.07, p=5.9e-10). These findings suggest that participants were able to implicitly infer the structure of the underlying graph.

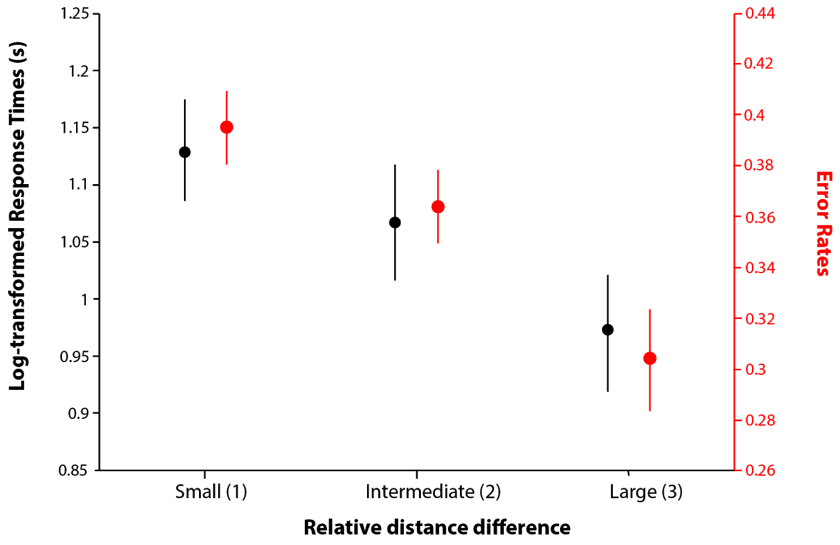

Figure 2. Participants are faster and more accurate at selecting the closer node when it was much closer than the alternative. Relative distance is discretized for display purposes.

After the judgement phase, we tested whether participants were also able to explicitly reconstruct this graph. On average, participants generated graphs that matched the true graph along two key metrics. First, these graphs successfully captured when two nodes were connected to one another (e.g., formed an edge in the graph; Figure 3a, Mean/Median Edge Accuracy = 84%/92%, SD = 4%). Their overall accuracy at identifying these edges was substantially higher than would be expected by chance (e.g., if participants had configured the nodes at random; Figure S1).

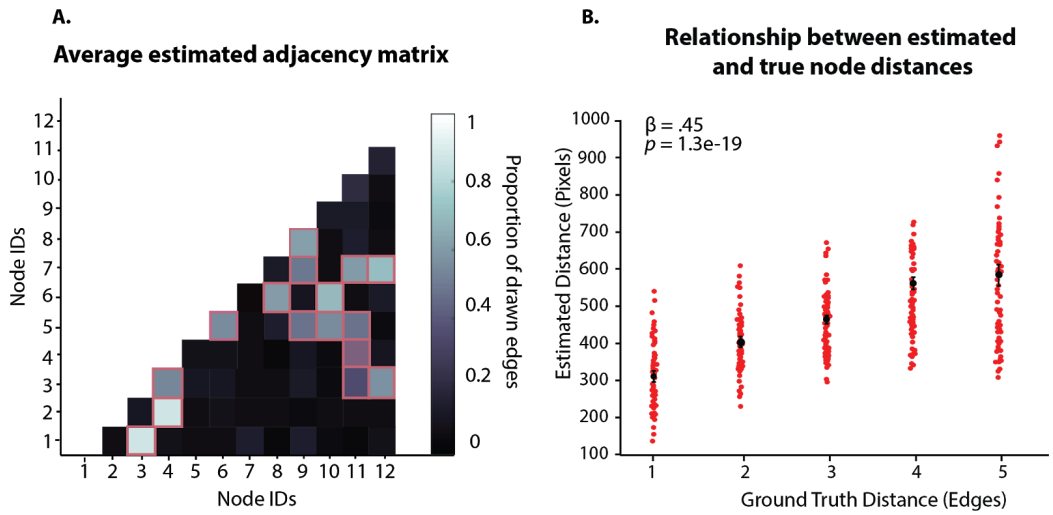

Figure 3. A) The proportion of drawn edges forming pairwise node connections in the graph. Lighter color indicates a higher proportion of connections. The squares outlined in orange correspond to the ground truth connections in the graph. Lighter color of fields outlined in orange indicates that the participants were more likely to draw edges between the nodes which are actually connected in the graph. **B)** Correlation between the ground truth distances (shortest paths in the graph), and pixel-based distances of recovered graphs.

Performance along this *adjacency* metric shows that participants were able to correctly identify the edges of the underlying graph, and therefore that they generally knew which nodes were connected to which other nodes. We also generated a second metric that examined the degree to which participants were able to also capture the relative *distances* between nodes in the graph (e.g., whether two connected nodes are close or far apart within the graph). In other words, when recreating the graph, to what degree was the Euclidean distance between objects placed on the screen (measured in pixels) representative of the true distance between the nodes in the underlying graph (measured in terms of the number of intervening edges/the shortest path). We found a significant correlation between these distance matrices (Figure 3B,

β=.45, t=12.4, p=1.3e-19, Mean(SD) Spearman r = .40(.25)). This relationship between estimated and true graph distance held even when controlling for the constructed adjacency matrix (β=.22, t=8.36, p=1.2e-11, Table S2), suggesting that this distance metric captured structure inference ability over and above participants' adjacency reports.

Our novel tasks thus provide evidence that people are able to infer the structure of a graph based on experiences with disjoint edges from the graph, with neither explicit instruction of the graph's existence, nor a task that encouraged them to learn this structure. We demonstrate this across six different measures: four from the relative distance judgment task (overall judgment accuracy, judgment accuracy by relative distance, search time by total distance, and search time by relative distance), and two measures from the graph reconstruction task (adjacency accuracy and the correlation between estimated (Euclidean) and actual (shortest path) graph distances). These six measures were highly correlated with one another across individuals (Figure S2C), suggesting that they reflect a common underlying dimension of individual differences in structure inference ability. Indeed, a principal component analysis (PCA) demonstrated that a single component (eigenvalue = 4.33) could capture 72% of the variance across these measures, and was the only substantive component that emerged in this analysis (all other eigenvalues < 0.85; Figure S2B). To examine the relationship between individual differences in structure inference ability and model-based planning, we therefore focused our analyses on variability in scores on this singular structure inference PC.

Better structure inference ability is associated with greater use of model-based planning

To measure individual differences in model-based planning, participants also performed a well-characterized assay of model-based planning, the two-step task (Figure 4, Daw et al., 2011; Gillan, Otto, Phelps, & Daw, 2015; Decker et al., 2016; Doll et al., 2015; Otto et al., 2013). In this task, participants make decisions at two stages that are connected by a probabilistic transition. To reach the best possible outcome on a given trial, participants must consider this underlying transition structure (e.g., engage in model-based planning). Previous studies show that choices in this task reflect a mixture of model-free and model-based forms of decision-making, indexing the degree to which participants choose actions based only on recent reward (*model-free*) or based additionally on a consideration of task structure (e.g., transition probabilities; *model-based*). We replicate this average pattern of behavior in our own data

(model-free index: β = .64, *t* = 14.38 , *p*= 3e-05, model-based index: β =.42, *t* = 9.17 , *p* = 1.5e-20; Figure 4C, Table S1).

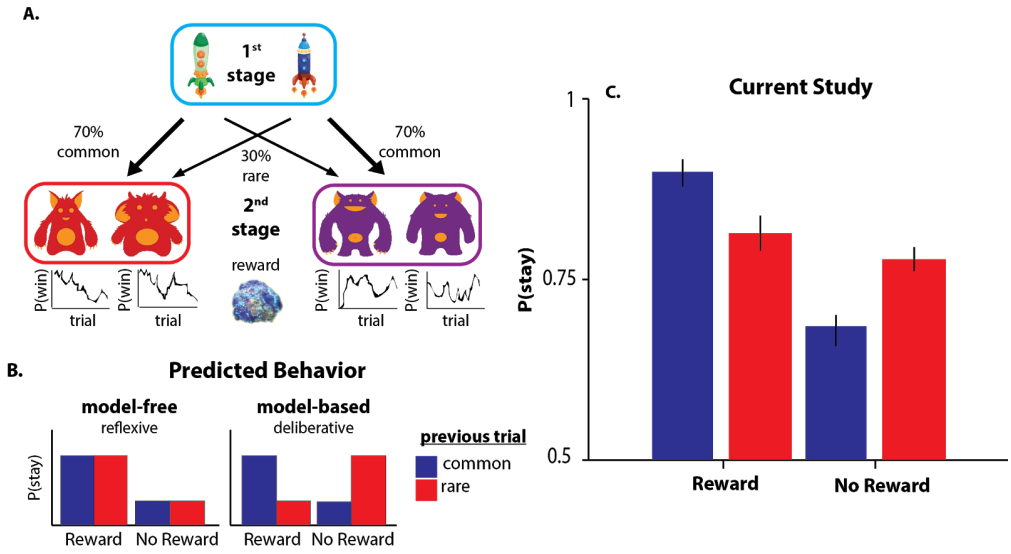

Figure 4. A) Two-step task design, adapted from Decker et al. (2016). The fixed structure of the probabilistic transitions from 1st stage states 2nd stage states enables the distinction of model-based and model-free choices by examining the influence of the previous trial on the subsequent first-stage choice. **B)** A model-free learner tends to repeat previously rewarded first-stage choices ("stay"), regardless of the transition type that led to the reward (a main effect of reward on subsequent first-stage choices). By contrast, a model-based learner exploits knowledge of the transition structure and will favor the first-stage action that is most likely to lead to the same state if rewarded and the action least likely to lead to the same state if not rewarded (a reward-by-transition interaction effect on subsequent first-stage choices). **C)** Consistent with previous literature, our results show that the participants exhibit a mixture of model-based and model-free choice strategies.

As in previous research, we also find that participants vary in their use of model-based planning on this task (e.g., as estimated by their model-based indices from the logistic regression). We predicted that these individual differences in model-based planning would be associated with individual differences in our index of structure inference ability, which we tested by examining whether there was a significant interaction between our structure inference PC and the model-based planning index in predicting first-stage choices (following previous work; Gillan et al, 2016). Consistent with our prediction, we found that participants who demonstrated better structure inference were also more likely to engage in model-based planning (β= .17, *t* = 3.97, *p* = .00008, Figure 5, Table S5). This correlation held when using an alternate estimate of model-based planning, based on an RL model of the two-step task (Figure 5, R^2 = .13, Spearman r = .33, *p* = .003; Table S4).

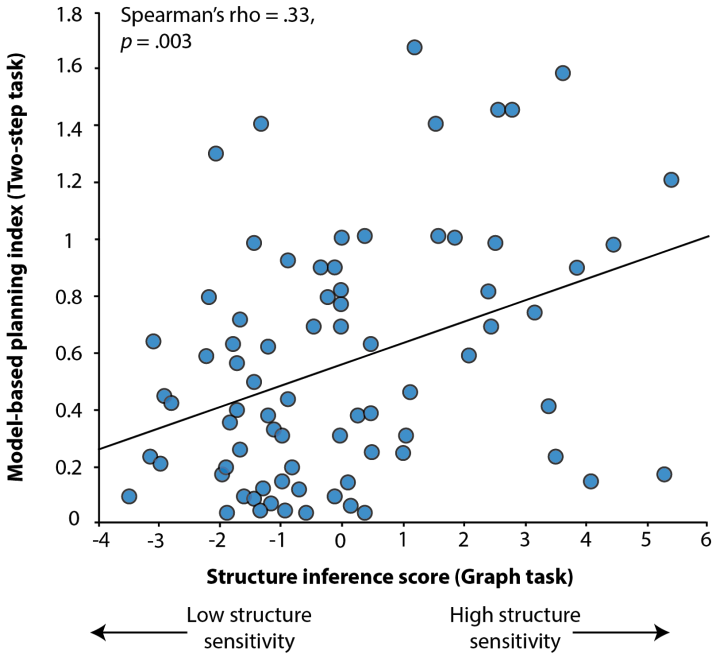

Figure 5. Latent factor capturing structure inference ability (PCA score) is positively correlated with model-based planning (model-based weights β^{MB}) across individuals.

Follow-up analyses confirmed that structural inference ability was specifically associated with model-based planning and not other aspects of two-step task performance. We controlled for individual differences in model-free strategy use and stay bias (perseveration), neither of which were associated with our structural inference index (model-free strategy use: β = -.04, t = -.17 , p = .86; perseveration: β = -.11, t = -.80 , p = .42, Table S3). We also tested whether the relationship between structural inference and model-based planning was mediated by individual differences in model-based *learning*. That is, while the structure of the two-step task is relatively simple (each of two nodes transitioning to two other nodes), and participants are made aware of the potential links between these nodes in advance (though not of the actual transition likelihoods), it could be that people who are generally worse at inferring graph structure are less likely to engage in model-based planning because they failed to learn the transition structure of the two-step task. Previous work has measured such individual differences in two-step task transition learning using a separate behavioral index: response times for Stage 2 decisions (following the transition from Stage 1). Overall, participants have been shown to respond slower in Stage 2 if they just experienced a rare transition rather than a common one (Decker et al, 2016), a finding that we replicate in our own data (β = -.01, t = 3.54 , p = .006). However, this

effect depends on having learned which transitions are rare and common, and therefore individual differences in the strength of this effect have been used to index individual differences in transition learning. Unlike model-based *planning*, this implicit index of model-based *learning* (post-rare transition slowing) was not significantly associated with structure inference ability (β =.06, t = .11, p = .57, Table S3). Moreover, controlling for model-based learning, the relationship between structure inference and model-based planning remained significant (β =.49, t = 2.19, p = .03, Table S3).

For completeness, we also tested whether the relationship between model-based planning and structure inference ability was specific to a subset of our structure inference metrics, but did not find that this was the case. Model-based planning was separately correlated with all of our structure inference metrics (all $|\beta|$>.12, all p<.007; Table S5).

Discussion

Goal-directed planning is critical for adaptive human behavior, providing the basis for long-term achievement across life domains. Successful goal-directed planning entails both (1) inferring the structure of one's environment (structure inference) and (2) deploying that structure to maximize reward (model-based decision-making), yet relatively little is known about how one's ability to do one of these relates to their ability to do the other. Combining six performance measures across a novel set of tasks, we characterized a dimension of structure inference ability, and showed that individual differences in this estimate of structure inference predicted individual differences in a well-characterized index of model-based planning (based on performance on the two-step task). We show that this association between structure inference and two-step task performance is specific to model-based planning rather than generalizing to other performance metrics, including perseveration, model-free decision-making, and a proxy for one's ability to learn transitions in that task. These results demonstrate that memory and decision-making share core cognitive substrates, with these connections across literatures potentially inform algorithmic models of knowledge-driven planning and decision-making.

Our work bridges previous research on structure learning and model-based planning, and fills an important gap in these earlier studies. In particular, previous research on model-based planning has only been able to examine how people learn about and navigate internal models

with limited nodes/connections (e.g., a single transition in the two-step task; Daw et al. 2011) and/or with transitions that are experienced sequentially in time (Bornstein & Daw, 2013; Doll et al., 2015). Using our novel tasks, we were able to examine how this inference process occurred for a more complex graph structure that participants learned based on disjoint experiences with individual node pairs. In doing so, we were able to establish that individuals are able to perform structure inference under such conditions and to exploit individual differences in their inferential abilities on these tasks to link the underlying cognitive processes to variability in model-based planning within simpler environments. Given this initial validation, our tasks hold promise for further bridging these lines of research to examine goal-directed navigation towards a reward (discussed below).

An important limitation of our study is that we do not have measures of structure inference taken prior to the judgment phase of the task (Phase 2). As a result, our measures of structure inference can reflect both (a) a participant's ability to infer the graph structure during the latent exposure phase (Phase 1) and (b) their ability to infer this graph structure in subsequent phases based on their initial exposure (possibly via directed search through these learned associations). Individual differences in either or both of these abilities could contribute to individual differences in model-based planning on the two-step task. While we were able to show that performance on the graph task was specifically related to model-based planning in the two-step task -- and not also related to a measure of structure (transition) learning on that task -- it is still possible that this correlation was partially driven by variability in structure learning ability not captured by the (indirect) measure available in the two-step task. A proper test of this question will require augmenting the existing task to allow a measure of online learning throughout our learning phase. Relatedly, our design also prohibits us from measuring the extent to which performance on the two-step task was influenced by performance on the structure inference tasks. Given our focus on individual differences, we fixed the order of these tasks to decrease heterogeneity across our sample, but future studies should counterbalance these tasks to estimate order effects and whether they vary systematically across participants.

While our study was able to provide an in-depth examination of how participants learn about the structure of the environment, it did so in the context of a fairly small structure (twelve nodes total) with deterministic transitions. This limits our ability to estimate how well our participants would be able to learn and traverse wider and more complex structures. Our graphs also

represent a fraction of the associative capacity of which people are capable, limiting the generalizability of our work to real-world structure learning and navigation. However, a number of the factors that enable individuals to learn structures on a much larger scale — such as temporal contiguity, spatial proximity, and semantic relations — are also factors that we sought to control in this experiment in order to minimize their role in learning and navigation. Future work can incorporate some of these additional elements to examine individual differences in learning larger and more complex graph structures (Schapiro et al. 2013; Diuk et al., 2014), including those with probabilistic and/or asymmetric transitions.

The observation that humans are able to rapidly learn complex graph structures without prior awareness raises several questions for further investigation. A particularly valuable direction for future research will be to investigate the neural mechanisms by which individuals infer the structure of our graph: Are these links inferred at encoding, during the learning task, or on-demand, at each trial during the judgement task? Evidence for both mechanisms was observed in a previous study that investigated transitive inferences across pairs of words in humans undergoing intracranial EEG (Reber et al. 2016). The authors reported that successful later inferences were predicted by hippocampal activity evident in both response-locked ERPs at test and stimulus-locked ERPs following encoding, providing support for multiple, hippocampally-centered mechanisms in the construction of inferences. This observation is consistent with a previously proposed distinction between prospective and retrospective integration in support of memory-guided decisions (Shohamy & Daw 2015; Ballard et al. 2019), and with recent work separately identifying a role for both encoding-time (Schapiro et al. 2013; Schapiro et al. 2016) and retrieval-time (Köster et al. 2018) computations in supporting similar inferential judgements. In both cases the judgements tested have been limited to a single step, though a more extensive capability was implied. Further work will be necessary to identify the relative contribution of each of these mechanisms to our task, in particular whether encoding and retrieval mechanisms are similarly useful in identifying extensive latent structure.

A critical factor in the ability to infer latent structure might be the effective use of hypotheses about the structure of the task. Specifically, humans and other animals are capable of *transfer learning*, or applying the schema learned in one instance of a task to another. Because our novel graph task relies on structures that can be clustered into families of resemblance and permuted to varying degrees, it is well-suited to measure the extent of the ability to transfer

across multiple instances. An important open question is to what degree these inferences are supported by general conceptual representations e.g. in PFC (Kumaran et al. 2012), structured basis representations in MTL cortex (Schapiro et al. 2016; Constantinescu et al. 2016; Behrens et al. 2018, or pre-generated associations cached in hippocampus proper (Collin et al. 2015). Because our task permits finely manipulating the relative information present in each kind of representation, it may be suitable for distinguishing involvement of each of these neural substrates.

Having established that the graph task measures a core aspect of model-based planning, future investigations could extend the task to incorporate planning for rewards directly, for instance by having participants learn about state-reward associations after (or in parallel with) learning the structure of state-state associations (Wimmer et al. 2012; Bornstein & Daw 2013). Examining how participants navigate this structure can provide important insight into how the structure of an internal model, and the relative distance between nodes in that model, modulates the utility of future rewards (Kurth-Nelson et al., 2015; Wimmer & Shohamy, 2012; Bornstein & Daw 2013) and the mental effort required to obtain those rewards (Kool & Botvinick., 2018, Shenhav et al., 2017). A similar design can also provide critical links to an expansive body of work on goal-directed navigation through space (Tolman, 1948; Tolman and Honzik, 1930; Tolman et al., 1946), including factors that influence individual differences in one's success in such navigation tasks (Maguire et al., 2000). In addition to highlighting common and divergent mechanisms across these domains, such research can also point toward potential training regimens that can be used to improve goal-directed planning, building on recent work in this area (Lieder et al., 2019). Collectively, these tasks can be used to compare and contrast neural mechanisms that underpin learning, navigation, and planning over latent structures across spatial and non-spatial domains.

Our work also provides important methodological and mechanistic insight for research on goal-directed decision-making across the lifespan, and between healthy and clinical populations. For instance, prior work has demonstrated that model-based planning gets worse with advanced aging (Eppinger et al., 2013) and negatively scales with behaviors indicating compulsive symptoms (Gillan et al., 2016). However, it remains unclear to what extent such impairments arise from deficits in inference (e.g., an inability to acquire and/or retain an internal model of the associative structure of one's environment) and/or deficits in the ability/motivation

to search that model to determine the best course of action (e.g., due to working memory demands and attendant mental effort costs). A better understanding of the mechanisms at the intersection of these processes will provide critical insight into the nature of goal-directed planning, and the factors that determine one's ability to achieve those goals.

Methods

Participants

We recruited 81 participants (38 female, Age range 18-27, Mean(SD) Age = 20(1.8)) from the Brown University participant pool. Participants received either course credit or monetary compensation for participating in the study. All participants provided informed consent in accordance with the policies of the Brown University Institutional Review Board. We excluded two participants with a high rate of perseverative responses in the two-step task (repeating the same response on more than 95% of the trials), and 2 participants due to the issues with data saving, resulting in the sample of total 77 participants included in the analyses.

Procedure

Two-Step Task

The two-step task is a sequential decision-making task, which enables assessment of dissociation between model-based and model-free choice strategies. In the task, participants made choices on two sequential stages, with the aim of obtaining a reward. In this version of the task (Decker et al, 2016), participants chose between two spaceships at stage one, which probabilistically transitioned to one of the two states (planets) at stage two (Figure 4A). In particular, each of the spaceships transitioned to one of the planets 70% of the time (common transition), and to the other planet 30% of the time (rare transition). These transition probabilities remained fixed throughout the task, and were taught to participants during training. At the second stage, participants encountered two aliens and chose one to solicit the space treasure/reward. The two aliens awarded treasure with independent probability, which shifted slowly over time according to a random Gaussian walk. Participants were instructed to earn as many pieces of treasure as possible. They had 3 seconds to make their choice on each stage. If they failed to make a response within the given time frame, the red 'x' appeared on top of the stimulus, and the trial terminated. Participants performed 40 practice trials, followed by 200 experimental trials.

The two-step task characterizes dissociable trial-by-trial adjustment of stage 1 choices, reflecting model-free and model-based choice strategies. On each trial, participants could choose between repeating the previous stage 1 choice and switching to the other spaceship. The model-free strategy predicts that the likelihood of staying or switching (repeating or changing the previous choice) on the first stage is informed by the outcome of the previous trial. The model-based strategy, on the other hand, predicts that the arbitration between staying and switching based on the observed outcome is modulated by the knowledge of transition type (common or rare) which on average led to that outcome over the course of trials. Thus, model-free reasoners choose to stay (repeat their prior stage 1 choice) if the outcome of that choice was rewarding on the previous trial, regardless of the transition type. On the other hand, model-based reasoners utilize the transition structure to select options that will most likely transition to the rewarding state (Figure 4B).

Graph task

Following the two-step task, participants performed a structure inference task designed to assess their ability to infer the latent structure based on the sequence of disjoint node pairs which, when reassembled, form the graph. They viewed a sequence of object pairs, each of which represented a pair of adjacent nodes drawn at random from an underlying undirected graph with 12 nodes and 16 edges (Figure 1A). Nodes in the graph were tagged by images of objects, which were randomly assigned for each participant. Each node-pair was presented for 1 second on the screen, after which the trial terminated and the next pair was presented.

Phase 1: Learning

In phase 1, participants passively viewed a sequence of object pairs, each displaying a pair of adjacent nodes. Node pairs were drawn at random, such that consecutive trials sampled adjacent node pairs from different locations on the graph (Figure 1B). Participants were not informed that there was an underlying structure, but were told that they would be tested on their memory of the pairs that had been presented. Each of the pairs was presented 44 times. In order to ensure that participants sustained attention throughout this phase, they were also asked to respond any time an object in the pair was rotated from its default position (which occurred on 10% of trials).

Phase 2: Relative distance judgment task

In phase 2, participants performed a relative distance judgment task, which required them to judge the relative distance between three randomly-selected nodes, unconstrained by edge relationship. In particular, participants viewed two nodes on either side of the screen and were asked to indicate which was closer to the reference node (shown centrally) based on the pairs they had seen in Phase 1 (Figure 1C). Participants were presented with 204 trials, and had an unlimited amount of time to make their choice.

Phase 3: Graph reconstruction

In phase 3, participants were shown all 12 nodes they had encountered in Phases 1 and 2 (Figure 1D). They were instructed to arrange the objects freely, by using their mouse to click on and move the object images on the screen. Once they positioned the nodes on the screen, participants were asked to connect them by clicking on pairs of images that they wished to group together. Participants had an unlimited amount of time to complete this stage of testing, and were allowed to make as many connections as they wished.

All of the tasks were programmed in Matlab (version 2016b; Natick, Massachusetts: The MathWorks Inc), using the Psychtoolbox 3 extension (Brainard, 1997; Pelli, 1997; Kleiner et al, 2007).

Analysis

Two-step task

Following previous work (Gillan, Otto, Phelps, & Daw, 2015; Daw et al, 2011; Decker et al, 2015), we quantified model-based behavior by performing a logistic mixed-effects regression analysis. We modeled participants' choice to stay (repeat the previous stage 1 choice) or switch on the current trial, as a function of (1) previous outcome (reward or no reward) (2) transition type (common or rare), and (3) a reward-by-transition type interaction (The model: *Stay ~ Previous Reward * Transition Type + (1+Previous Reward * Transition Type | Participant*). The main effect of the reward in the model is an index of model-free behavior, quantifying participant's choices as a function of recent outcome. The reward-by-transition interaction term serves as an index of model-based behavior, as it captures how much participants' choices were affected by the recent outcome, modulated by the knowledge of the transition structure. Therefore, variability in the interaction term demonstrates individual differences in how much

participants relied on model-based reasoning in the task. The regression included maximal random slopes and intercepts for each participant.

Reinforcement Learning Model

Participants' full trial-by-trial choice sequence in the task can also be fit with a computational reinforcement-learning model that gauges the degree to which participants choices are better described by a model-based or model-free reinforcement learning algorithm. Indices of model-based learning in the two-step RL task were derived via Bayesian estimation using a variant of the computational model introduced in (Daw et al., 2011). The model assumes choice behavior arises as a combination of model-free and model-based reinforcement learning. Each trial t begins with a first-stage choice $c_{1,t}$ followed by a transition to a second state s_t where the participant makes a 2nd stage choice $c_{2,t}$ and receives reward r_t. Upon receipt of reward r_t, the expected value of the chosen 2nd stage action (the left vs. the right alien) $Q_t^{s2}(s_t, c_{2,t})$ is updated in light of the reward received.

According to the model, the decision-maker uses a learned value function over states and choices $Q^{s2}(s,c)$ to makes second-stage choices. On each trial, the value estimate for the chosen action is adjusted towards the reward received using a simple delta rule, $Q_{t+1}^{s2}(s_t, c_{2,t}) = (1-\alpha)Q_t^{s2}(s_t, c_{2,t}) + r_t$, where α is a free learning rate parameter that dictates the extent to which value estimates are updated towards the received outcome on each trial. Unlike the standard delta rule ($Q_{t+1}^{s2}(s_t,a) = (1-\alpha)Q_t^{s2}(s_t,a) + \alpha r_{t.}$), in this equation and in similar references throughout, the learning rate α is omitted from the latter term. Effectively, this reformulation rescales the magnitudes of the rewards by a factor of $1/\alpha$ and the corresponding weighting (e.g., temperature) parameters β by α. The probability of choosing a particular 2nd stage action $c_{2,t}$ in state s_t is approximated by a logistic softmax, $P\left(c_{2,t} = c\right) \propto \exp\left(\beta^{s2} Q_t^{s2}(s_t, c_2)\right)$ with free inverse temperature parameter β^{s2} normalized over both options c_2.

First-stage choices are modeled as a product of both model-free and model-based value predictions. The model-based value of each 1st stage choice is dictated by the learned value of the corresponding 2nd stage state, maximized over the two actions: $Q_t^{MB}(c_1) = max(Q_t^{s2}(s,c_2))$,

where s is the second-stage state predominantly produced by first-stage choice c_1. Model-free values are governed by two learning rules, TD(0) and TD(1), each of which updates according to a delta rule towards a different target. Whereas earlier models posit a single model-free choice weight β_{MF} and use an eligibility trace parameter $\lambda \in (0, 1)$ to control the relative contributions of TD(0) and TD(1) learning, here, as in recent work by Gillan et al., (2016), model-free valuation is split into its component TD(0) and TD(1) stages, each with separate sets of weights and Q values. TD(0) backs-up the value of the stage-1 choice on the most recent trial $Q_{t+1}^{MF0}\left(c_{1,t}\right)$ with the value of the state-action pair that immediately (e.g., lag-0) followed it: $Q_{t+1}^{MF0}\left(c_{1,t}\right) = (1-\alpha)\, Q_t^{MF0}\left(c_{1,t}\right) + Q_t^{s2}\left(s_t,\, c_{2,t}\right)$. TD(1), on the other hand, backs up its value estimate $Q_{t+1}^{MF1}\left(c_{1,t}\right)$ by looking an additional step ahead (e.g., lag-1) at the reward received at the end of the trial: $Q_{t+1}^{MF1}(c_{1,t}) = (1-\alpha) Q_t^{MF1}(c_{1,t}) + r_t$. Ultimately, Stage-1 choice probabilities are given by a logistic softmax, where the contribution of each value estimate is weighted by its own model free temperature parameter:

$$P\left(c_{1,t} = c\right) \propto \exp\left(\beta^{MB} Q_t^{MB}(c) + \beta^{MF0} Q_t^{MF0}(c) + \beta^{MF1} Q_t^{MF1}(c)\right).$$

At the end of each trial, the value estimates for all unchosen actions and unvisited states are multiplicatively re-weighted by a free discount parameter γ [0,1]. This conventional parameterization reflects the assumption that value estimates decay exponentially at a rate of $1-\gamma$ over successive trials (Ito & Doya, 2009; Hunter et al., 2018). The temporal decay of value is widely endorsed by normative and empirical research on reinforcement learning (Sutton & Barto, 1998; Ito & Doya, 2009), and is further motivated endogenously by the perseverative nature of choice behavior in this task. Earlier models of behavior in this task have operationalized choice perseveration using a "stickiness" parameter, which is implemented as a recency bonus or subjective "bump" in the value of whichever first stage action was chosen on the most recent trial (irrespective of reward) (Daw et al., 2011). Including a decay parameter also accounts for the fact that people tend to 'stay' (repeat) their previous 1st stage choice: as γ approaches 0, the value of the unchosen actions decreases relative to the value of the chosen action on the next trial regardless of whether or not a reward was received (Hunter et al., 2018). In total the model has six free parameters: four weights $(\beta^{s2},\ \beta^{MB}, \beta^{MF0}, \beta^{MF1})$, a learning rate α

, and a decay rate γ. The six free parameters of the model $\left(\beta^{s2},\ \beta^{MB}, \beta^{MF0}, \beta^{MF1}, \alpha, \gamma\right)$ were estimated by maximizing the likelihood of each individual's sequence of choices. Numerical optimization was used to find maximum likelihood estimates of the free parameters, with 10 random initializations to help avoid local optima.

Graph task

We first assessed whether choice difficulty predicts accuracy and RTs in the relative distance judgment task. We defined choice difficulty as the absolute difference between exemplars' distances from the reference node, with distance defined as the shortest path length. The greater the distance difference between options and the reference, the closer one option was to the reference relative to the other, thus making the discrimination easier. In addition, we tested whether participants' response times also scaled with the total distance of option nodes from the reference node (e.g. depth search). As with the regression approach in the two-step task, we computed random effects for all subjects in these regressions as well.

We quantified how similar each participant's recovered graph was to the true graph using two metrics. First, we looked at the average similarity between the adjacency matrix of the ground truth and the recovered graph (Figure 3A). Second, we tested whether we can predict the pairwise ground truth distance (the shortest path) with the Euclidean distance between any two node placements (Figure 3B).

Relationship between the performance on the graph task and the two-step task

The overarching aim of this study was to examine whether there is a relationship between the deployment of model-based planning (quantified by the interaction term from the two step task, and the model-based weights from the computational model), and structure inference ability (determined by the measures from the graph task). To answer this question, we looked at the relationship between the model-based indices, and six main measures from the graph task (1) overall judgment accuracy, random effect estimates of task difficulty on 2) response times and 3) accuracy, 4) effect of total distance on the search time, 5) percentage of correctly identified edges, 6) Fisher transformed pixel-node distance correlations). We added these separate metrics from the graph task as covariates in the 1-back *stay ~ reward * transition* logistic regression model, in order to estimate direct association between model-based planning and

these between-subject factors. We report results from independent models, with each z-scored covariate *Z* entered separately: *stay ~ reward * transition type * Z + (1+reward*transition type|Participant)*. In addition, using principal component analysis (PCA) we collapsed the six graph measures into a single latent factor which captures variability in structure inference performance. We then performed the same analysis described above, where we entered participants' PCA scores as z-scored covariates in the two-step task regression model. To validate consistency between our two approaches to extracting subject-level model-based planning indices from the two step task (the interaction term from the logistic regression, and the model-based weighting parameter (β_{MB}) from the computational model), we also tested the association between participants' PCA scores and β_{MB} by performing a robust linear regression and Spearman correlation.

References

Ballard, I., Wagner, A. and McClure, S. (2019). Hippocampal pattern separation supports reinforcement learning. *Nature Communications*, 10(1).

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology, 37(4-5), 407-419.

Balleine, B. W., & O'Doherty, J. (2009). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology, 35, 48e69.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. Neuron, 100(2), 490-509.

Bornstein, A. M., & Daw, N. D. (2013). Cortical and Hippocampal Correlates of Deliberation during Model-Based Decisions for Rewards in Humans. PLoS Computational Biology, 9(12). doi:10.1371/journal.pcbi.1003387

Brainard, D. H. (1997) The Psychophysics Toolbox, *Spatial Vision 10*:433-436.

Collins, A. & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* 8, 240-247.

Collin, S., Milivojevic, B. and Doeller, C. (2015). Memory hierarchies map onto the hippocampal long axis in humans. *Nature Neuroscience*, 18(11), pp.1562-1564.

Constantinescu, A.O., O'Reilly, J.X., Behrens, T.E.J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science* 352(6292):1464–1468..

Daw, N., Gershman, S., Seymour, B., Dayan, P., & Dolan, R. (2011). Model-Based Influences

on Humans Choices and Striatal Prediction Errors. *Neuron*, 69(6), 1204-1215. doi:10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience, 8, 1704e1711.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From Creatures of Habit to Goal-Directed Learners. *Psychological Science*, 27(6), 848-858. doi:10.1177/0956797616639301

Diuk, C., Schapiro, A., Cordova, N., Fernandes, R.J., Niv, Y., & Botvinick, M. (2014). Divide and Conquer: Hierarchical Reinforcement Learning and Task Decomposition in Humans. *Computational and Robotic Models of the Hierarchical Organization of Behavior*. 271-291. 10.1007/978-3-642-39875-9_12.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. Neuron, 80, 312e325

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. Nature Neuroscience, 18, 767e772

Doll, B. B., Shohamy, D., & Daw, N. D. (2015). Multiple memory systems as substrates for multiple decision systems. *Neurobiology of Learning and Memory, 117*, 4-13. doi:10.1016/j.nlm.2014.04.014

Eppinger, B., Walter, M., Heekeren, H. and Li, S. (2013). Of goals and habits: age-related and individual differences in goal-directed decision-making. *Frontiers in Neuroscience*, 7.

FeldmanHall, O & Shenhav, A. (2019). Resolving uncertainty in a social world. Nature Human Behaviour 3: 426-435.

Fermin, A., Yoshida, T., Ito, M., Yoshimoto, J., & Doya, K. (2010). Evidence for model-based action planning in a sequential finger movement task. Journal of motor behavior, 42(6), 371-379.

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience, 15(3),* 523-536. doi:10.3758/s13415-015-0347-6

Hunter, L.E., Bornstein, A.M., Hartley, C.A. (2019). A common deliberative process underlies model-based planning and patient intertemporal choice. bioRxiv. doi:10.1101/499707

Ito, M. and Doya, K. (2009). Validation of Decision-Making Models and Analysis of Decision Variables in the Rat Basal Ganglia. *Journal of Neuroscience*, 29(31), pp.9861-9874.

Johnson A., Redish A.D (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J. Neurosci. 2007;27:12176–12189

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. Cognitive science, 20(2), 137-194.

Kleiner, M., Brainard, D., Pelli, D. (2007). "What's new in Psychtoolbox-3?" Perception 36 ECVP Abstract Supplement.

Kool, W. and Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, 2(12), pp.899-908.

Köster, R., Chadwick, M., Chen, Y., Berron, D., Banino, A., Düzel, E., Hassabis, D. and Kumaran, D. (2018). Big-Loop Recurrence within the Hippocampal System Supports Integration of Information across Episodes. *Neuron*, 99(6), pp.1342-1354.e6.

Kumaran, D., Melo, H. and Duzel, E. (2012). The Emergence and Representation of Knowledge about Social and Nonsocial Hierarchies. *Neuron*, 76(3), pp.653-666.

Kurth-Nelson, Z., Barnes, G., Sejdinovic, D., Dolan, R. and Dayan, P. (2015). Temporal structure in associative retrieval. *eLife*, 4.

Lieder, F., Chen, O., Krueger, P.M., & Griffiths, T.L., (2019). Cognitive Prostheses for Goal Achievement. 10.13140/RG.2.2.16279.06564/1.

Maguire, E. A., Woollett, K., & Spiers, H. J. (2006). London taxi drivers and bus drivers: A structural MRI and neuropsychological analysis. *Hippocampus, 16*(12), 1091-1101. doi:10.1002/hipo.20233

MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.

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

Otto, A.R., Raio, C.M., Chiang, A., Phelps, E.A., & Daw, N.D. (2013). Working-Memory Capacity Protects Model-Based Decision-Making from Stress. *Proceedings of the National Academy of Sciences, 110(52),* 20941-20946.

Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2015). Cognitive Control Predicts Use of Model-based Reinforcement Learning. *Journal of Cognitive Neuroscience*, 27(2), 319-333. doi:10.1162/jocn_a_00709

Parkinson, C., Kleinbaum, A., & Wheatley, T. (2017). Spontaneous Neural Encoding of Social Network Position. Nature Human Behavior, 1, 1-7.

Pelli, D. G. (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies, *Spatial Vision 10*:437-442.

Reber, T., Do Lam, A., Axmacher, N., Elger, C., Helmstaedter, C., Henke, K. and Fell, J. (2015). Intracranial EEG correlates of implicit relational inference within the hippocampus. *Hippocampus*, 26(1), pp.54-66.

Schapiro, A., Kustner, L., & Turk-Browne, N. (2012). Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. *Current Biology, 22*(17), 1622-1627. doi:10.1016/j.cub.2012.06.056

Schapiro, A., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2016). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1711), 20160049. doi:10.1098/rstb.2016.0049

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T., Cohen, J. and Botvinick, M. (2017). Toward a Rational and Mechanistic Account of Mental Effort. *Annual Review of Neuroscience*, 40(1), pp.99-124.

Shenhav, A., Rand, D. and Greene, J. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, 141(3), pp.423-428.

Shohamy, D. and Daw, N. (2015). Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences*, 5, pp.85-90.

Shohamy, D., & Turk-Browne, N. B. (2013). Mechanisms for widespread hippocampal involvement in cognition. Journal of Experimental Psychology: General, 142, 1159-1170.

Sutton, R. and Barto, A. (1998). Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks*, 9(5), pp.1054-1054.

Tamir, D.I., Thornton, M.A. (2018). Modeling the predictive social mind. Trends in Cognitive Science, 22(3), 201-212.

Tolman, E. C. (1948). Cognitive maps in rats and men. Psychological Review, 55(4), 189-208.

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, E.C., Ritchie, B.F., Kalish, D.(1946). Studies in spatial learning. I. Orientation and the short-cut. Journal of Experimental Psychology. 36:13–24

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

Wimmer, G. and Shohamy, D. (2012). Preference by Association: How Memory Mechanisms in the Hippocampus Bias Decisions. *Science*, 338(6104), pp.270-273.

Wimmer, G., Daw, N. and Shohamy, D. (2012). Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience*, 35(7), pp.1092-1104.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. European journal of neuroscience, 19(1), 181-189.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action–outcome learning in instrumental conditioning. European Journal of Neuroscience, 22(2), 505-512.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2006). Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning. Behavioural brain research, 166(2), 189-196.

Supplemental Figures

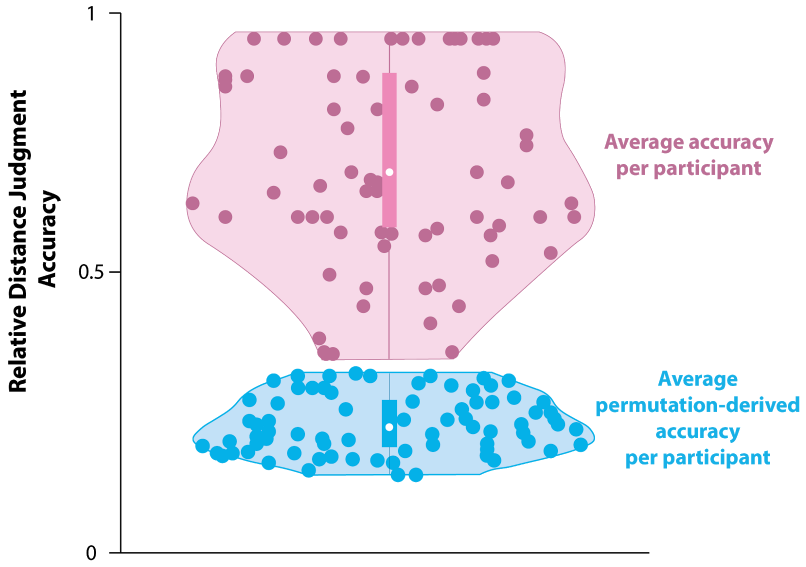

Figure S1. Distribution of (1) participant accuracy on graph reconstruction task (the percentage of correctly identified edges) and (2) chance level (permutation-derived) accuracy for each subject. We estimated chance accuracy for each subject by randomly shuffling the graph 1000 times (such that connections between nodes differ from those in the ground truth graph), evaluating the accuracy of connected edges based on the adjacency matrix of the shuffled graph and averaging accuracy values in each subject. The edge connections based on the shuffled graph are a proxy of how accurate subjects would be if they were guessing at random while drawing the edges between nodes. Our results show that participants' edge connection accuracy evaluated based on the ground truth adjacency matrix is significantly greater than the chance accuracy based on the randomly shuffled graphs, confirming that on average participants were not randomly configuring the edges.

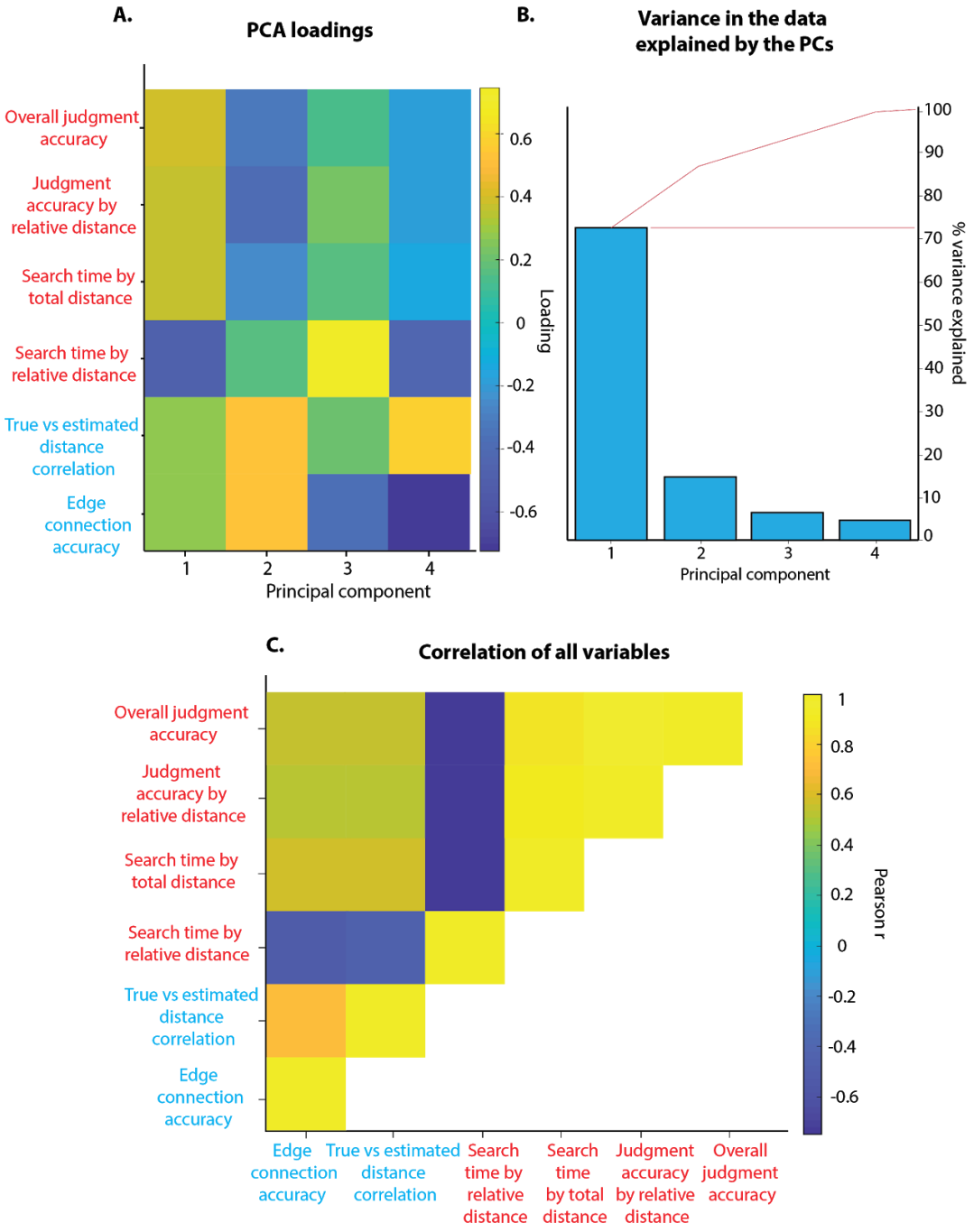

Figure S2. PCA results and pairwise correlations of all graph-task measures. Plot A shows the PC loadings on all 6 graph-task measures (red = relative distance judgment task measures; blue = graph reconstruction measures). First component loads on all 6 measures, whereas the second component is selective for graph reconstruction measures. Eigenvalue of the first component is 4.33, and the second component is 0.85. Plot 4B shows the percentage of variance captured by different components. The first component captures the majority of variance (72%). Plot 4C shows the pairwise correlations between graph-task measures.

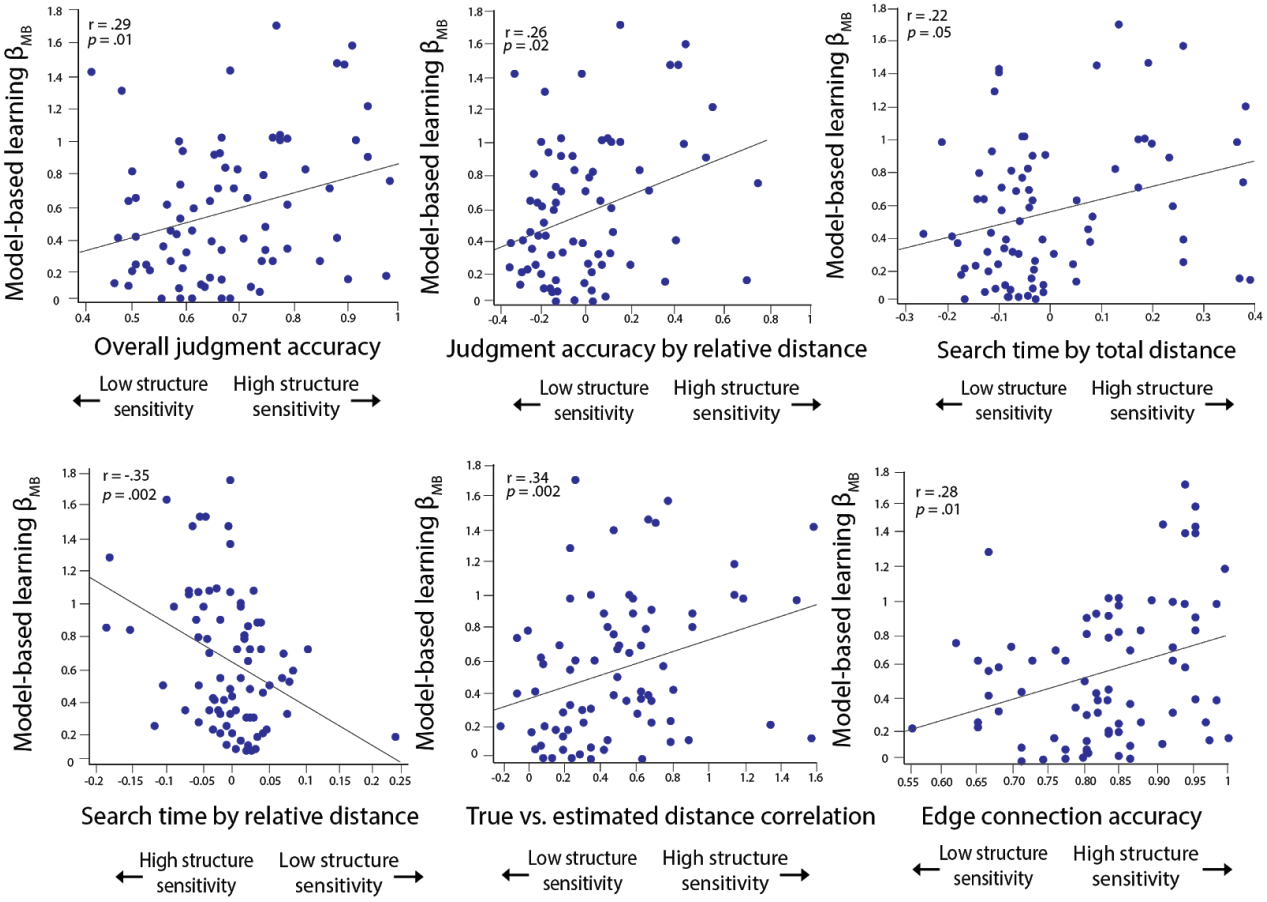

Figure S3. ***Model-based learning is associated with individual structure inference measures.*** The y-axis corresponds to the fit value of the model-based weighting parameter β_{MB}. The x-axis in the above figures plots the following: overall judgment accuracy, judgment accuracy by relative distance, search time by total distance, search time by relative distance, true vs. estimated distance correlation, edge connection accuracy.

Supplemental Tables

	β(SE)	*T*	DF	*p*
Transition type	.10(.02)	3.9	73	.0001***
Reward	.64(.04)	14.38	73	2.9e-05***
Transition type * Reward	.42(.04)	9.17	73	1.5e-20***

Table S1. ***Modeling 1st-stage choices in the RL task as a function of model-free and model-based learning.*** Model statistics refer to the coefficients of the fixed main effect of reward, transition type and the reward-by-transition type interaction from the following model: *Stay ~ Reward × Transition type + (1 + Reward × Transition type | Participant)*. Here (SE) indicates the standard error of the mean (*p<.05; **p<.01; ***p<.001). Participants exhibit a mixture of model-free and model-based strategies, as shown by the significant main effect of reward and a reward-by-transition type interaction.

	β(SE)	*T*	DF	
Estimated shortest path	.42(.05)	7.81	74	5.63e-11***
Euclidean/ distance	.22(.02)	8.36	74	1.22e-11***

Table S2. Predicting ground truth distance with Euclidean distance, controlling for reported shortest path. The Euclidean distance (estimated distance) predicts the ground truth distance over and above the reported shortest path. This suggests that the estimated distance model was not predictive of the ground truth simply as a function of edge connections participants drew during the reconstruction. DF here refers to Satterthwaite degrees of freedom approximation. (*p<.05; **p<.01; ***p<.001).

	***β*(SE)**	***T***	**DF**	***p***
Model-based term random effects	.49(.22)	2.19	72	.03**
Perseveration percentage	-.11(.14)	-.80	72	.42
Model-free term random effects	-.04(.23)	-.17	72	.86
Post-rare transition slowing random effects	.06(.11)	.55	72	.57

Table S3. Robust linear regression model predicting PC scores (latent component indexing structure inference) using different indices from the two-step task (percentage of perseverative response, model-free random effects, model-based random effects and post-rare transition slowing). Index of model-based planning is selectively significantly associated with the measure of structure inference. The beta coefficients here are estimated effect coefficients, SE is standard error of the mean. DF refers to error degrees of freedom. Positive terms indicate positive association with the structure inference measure (*p<.05; **p<.01; ***p<.001).

	β(SE)	***T***	**DF**	***p***
Latent structure learning factor	.16 (.04)	3.35	75	.001**

Table S4. ***Model-based weights from the computational model.*** The results from the robust linear regression model (*RL Model-based weights ~ 1 + structure inference score*)**.** The predictor in the model was z-scored. The results show that high PC scores (indexing high structure inference performance) predict increased model-based planning. (*p<.05; **p<.01; ***p<.001).

Structure inference measure	*β (SE)*	*T*	DF	*p*
Overall distance judgement accuracy	.14 (.05)	3.16	72	.001**
Judgment accuracy by relative distance	.16 (.04)	3.60	72	.0003***
Search time by total distance	.13(.04)	2.95	72	.003**
Search time by relative distance	-.16 (.04)	-3.69	72	.0002***
True vs estimated distance correlation	.18 (.04)	3.98	72	.00008***
Edge connection accuracy	.12 (.04)	2.67	72	.007**
Latent structure learning factor	.17(.04)	3.97	72	.00008***

Table S5*. Modeling 1st-stage choices in the RL task as a function of structure inference ability and model-based planning.* Each row reflects the results from an independent analysis where each covariate (z-transformed) was entered as Z in the following model: *Stay ~ Reward × Transition × Z + (1 + Reward × Transition | Participant)*. Model statistics refer to the coefficient of the fixed-effects interaction: *Reward × Transition × Z*. Positive values indicate an association with increased model-based planning Covariates with positive values are associated with increased model-based learning (except for the search time by relative distance effect). Here (SE) indicates the standard error of the mean. (*p<.05; **p<.01; ***p<.001).