Dynamic preferences account for inter-animal variability during the continual learning of a cognitive task

David B. Kastner[1,2], Eric A. Miller[2], Zhounan Yang[1], Demetris K. Roumis[2], Daniel F. Liu[2], Loren M. Frank[2,3*], Peter Dayan[4*]

[1]*Department of Psychiatry, University of California, San Francisco, CA 94143, USA*

[2]*Kavli Institute for Fundamental Neuroscience and Department of Physiology, University of California, San Francisco, CA 94158, USA*

[3]*Howard Hughes Medical Institute*

[4]*Max Planck Institute for Biological Cybernetics, Tübingen 72076, Germany*

[*]*These authors contributed equally to this work*

Lead contact: David B. Kastner: david.kastner2@ucsf.edu

Summary

In novel situations, behavior necessarily reduces to latent biases. How these biases interact with new experiences to enable subsequent behavior remains poorly understood. We exposed rats to a family of spatial alternation contingencies and developed a series of reinforcement learning agents to describe the behavior. The performance of these agents shows that accurately describing the learning of individual animals requires accounting for their individual dynamic preferences as well as general, shared, cognitive processes. Agents that include only memory of past choice do not account for the behavior. Adding an explicit representation of biases allows agents to perform the task as rapidly as the rats, to accurately predict critical facets of their behavior on which it was not fitted, and to capture individual differences quantitatively. Our results illustrate the value of making explicit models of learning and highlight the importance of considering the initial state of each animal in understanding behavior.

Introduction

Learning requires storing new information and using it to guide behavior. Standard accounts of learning often study this process without regard to an animal's proclivities, in part because it is difficult to account for these pre-existing biases quantitatively. Nonetheless, in cases where specific biases can be identified in individuals (such as sign- or goal-tracking (Flagel et al., 2011)), accounting for these biases provides important predictive information about subsequent behavior (Flagel et al., 2010; Stead et al., 2006).

Ideally one would have a simple quantitative account of the initial state of each subject upon entering into a new situation, and this account could then be augmented by a description of how that state evolved as the subject incorporated new experiences into its knowledge base and used stored memories to guide its decisions. Reinforcement learning (RL) models provide a path forward toward that goal, particularly in the context of behaviors that center on rewards and punishments (Daw, 2011; Gershman and Daw, 2017), but RL is most often applied to asymptotic performance rather than the initial learning of tasks. Consequently, we lack a systematic approach to describe and incorporate the initial state of subjects and their entire course of learning. Furthermore, animals seamlessly learn different tasks over many different timescales, a characteristic with which machine learning and artificial intelligence are just starting to grapple (Kirkpatrick et al., 2017; Zenke et al., 2017). Yet it is rather rare for behavioral experiments to systemize, let alone study, the way that a task is learned from scratch. It is thus not surprising that models of behavior often fail to capture the way that animals learn tasks, measured by key characteristics such as the speed of acquisition.

Capturing the initial state and the entire course of learning is difficult for experimental, empirical, statistical and computational reasons. Experimentally, animals are typically exposed to behavioral tasks without prior experiences that would aid in identifying (or controlling) pre-existing biases. Empirically, at least until recently (International Brain Laboratory, 2017), it has been common to shape the behavior of each subject in an individualized and heterogeneous manner, a natural result of the focus on asymptotic performance of tasks rather than the acquisition of those tasks. Statistically, each subject only provides one set of data points about their entire learning trajectory, implying that one needs a large number of subjects, or very

powerful inductive biases. Computationally, standard RL agents typically learn either far slower or far faster than animals. In general, agents learn more slowly, since they do not capture the extensive knowledge about the world that animals apparently possess. However, particularly model-based agents (Gershman and Niv, 2010; Lake et al., 2017), can be constructed to efficiently absorb the relevant information (or be provided with an advantageously restricted set of inputs), and thereby (unfairly) gain information from the outset as to which are the critical features of the environment. This allows them to learn more quickly than the subjects. However, the relevance of that rapid acquisition to the learning that actually goes on in the subjects is unclear.

These problems led us to develop a new experimental and a new modeling approach to understand pre-existing biases and the dynamics of learning. We began with a specially designed rich and complex spatial alternation paradigm (Singer and Frank, 2009; Singer et al., 2010) with an initial set of reward contingencies that was designed to allow us to identify pre-existing biases and substrates for dynamic preferences. To minimize the issues of tailored shaping for the learning of each individual animal, we developed an automated behavioral system that administered the series of contingencies to the rats. We then undertook the task of modeling much of the course of the behavioral performance using a family of models with progressively increasing complexity. To help with the statistical issues, we exploited a key feature of the paradigm—that it evolves in discrete stages, each of which poses specific challenges—thus providing repeated windows into aspects of early acquisition.

The final model can capture many of the key features of learning and provides insights into what the animals know at each point in the task. The model not only fits the aggregate performance of the animals, but it also captures aspects of the variability among the individual rats, including differences in initial biases that strongly predict subsequent learning performance. As such, the model provides a platform from which individual variability (Galsworthy et al., 2002; Matzel et al., 2003) can be quantified and thus studied.

Results

Automated system to enable systematic control and measurement of behavior

One of our goals was to standardize behavioral training and to reduce potential effects of experimenter-subject interactions on learning (Sorge et al., 2014). We therefore developed an automated behavioral system (Brunton et al., 2013; Poddar et al., 2013; Rivalan et al., 2017) that required minimal animal handling: once animals were placed in the apparatus, no further experimenter contact was necessary until the end of the daily behavior. This system also enables the measurement of behavior across many animals throughout the entire course of learning and performance of the task. This apparatus itself contains four parts: 1) a six-armed track with reward wells at the end of each arm; 2) four rest boxes, each with a reward well; 3) corridors connecting the rest boxes to the track; and 4) doors to gate the pathway on and off the track for each rest box (Fig 1A).

The sequence of operations of the track for the set of behaviors are: 1) the doors open to clear the path from a single rest box to the track. Concurrently, the lights linked to all of the

reward wells on the track turn on (Fig S1A). 2) On the first break of a track reward well beam (Fig S1A) following the opening of the doors, the door to the track closes, thus starting the session of that animal. The animal then has a fixed maximum number of trials for its session, and the session ends when either that maximum has been reached, or following a time limit of 30 minutes. All but one of the animals never reached the time limit (see methods). 3) Upon breaking the beam at the reward well at the last trial of the session, all of the reward well lights turn off, and the doors reopen to allow for passage back to the appropriate rest box. Concurrently, the light to the reward well in that rest box turns on. 4) Upon breaking the beam of the rest box reward well, the doors to the track close and the well delivers reward. The light of the rest box reward well turns off after reward delivery. 5) The doors to the track for the rest box for the next subject open, and the process repeats itself.

Rats learn a series of spatial alternation contingencies

The six arms of the track allow for the learning of multiple spatial contingencies. We used a spatial alternation task, which is commonly used to study learning, memory and decision making, both during normal function and in the context of models of neuropsychiatric disease (Dudchenko, 2004; Frank et al., 2004; Karlsson and Frank, 2008; Sigurdsson et al., 2010). The task requires an animal to learn, through trial and error, to visit goal location in a predetermined order which depends on past actions and associated outcomes. Such a task has traditionally been utilized to explore hippocampal function. However, even substantial manipulations of the hippocampus do not render subjects totally incapable of performing the task (Fernandez-Ruiz et al., 2019; Jadhav et al., 2012; Kim and Frank, 2009), indicating that other brain regions and memory systems are at least partially independently competent (Packard and Goodman, 2013; Packard and McGaugh, 1996).

One of our key goals was to identify pre-existing biases in choice behavior. That goal motivated the inclusion of an initial exploration period: prior to the rats beginning the spatial alternations task, they had 14 – 16 sessions (362 – 425 trials) of exploration on the track wherein they were rewarded at any arm that they visited as long as it was not a repeat visit to the immediately preceding arm. In addition to providing information about biases in the absence of challenging reward contingencies, this exploration provided the rats with the opportunity: 1) to get used to entering and exiting their rest boxes when the doors to the track opened, and 2) to explore the space of the track. This functionally separated the learning of the environment from the learning of the task.

Our analyses of the arm visits during these exploration sessions revealed multiple types of bias. Individual rats showed preferences towards visiting specific arms (Fig 1B&E). The rats also had a large propensity to transition to neighboring arms (Fig 1C&E). And finally, they exhibited directional inertia, whereby they continued to move in the same direction on the current trial as they did in the immediately preceding trial (Fig 1D&E). We incorporated these biases sequentially into our models to determine whether they were important contributors to, or modulators of, the subsequent learning processes.

Following this initial exploratory period, and without any external signal to indicate a change, rats were sequentially exposed to different spatial alternation contingencies. In each contingency, only three arms had the potential to deliver reward. Reward would be delivered

within a given contingency if the rat alternated between the outer arms after every visit to the center arm. For instance, if the contingency were at arms 2-3-4, then to get reward the animal would have to follow the sequence 3-4-3-2-3 etc.

The algorithm that governs this task defines two trial types, inbound and outbound. If the rat was at any arm other than the center arm, the way to get reward was to go to the center arm (arm 3 in this example); we will refer to these trials as inbound trials. Any such trial on which the rat failed to go into the center arm (and any error during the exploratory contingency) is called an inbound error. Once the rat had visited the center arm, then the only way it could get reward is to visit the less recently visited outer arm (i.e. in this example if before going to the center arm 3 the rat came from arm 4, then it would have to go to arm 2 next to get a reward). We will refer to these trials as outbound trials, and the corresponding error as outbound errors. Note that apparent alternative policies for solving the task, for instance by alternating rightward and leftward sequences (3-4; then 3-2; then 3-4; etc.), need not necessarily recover from errors in quite the way that the governing algorithm rewards maximally.

The rats learned a series of six different spatial alternation contingencies (Fig 1F), chosen to present increasing challenges. The transition from the first (2-3-4) to the second (1-2-3) contingency was designed to be relatively easy, since performing 2-3-4 would allow it to find the central arm of the new contingency readily. Finding this arm is critical to gaining consistent reward. The transition from the second (1-2-3) to the third (3-4-5) contingency was designed to be harder since the central arm (4) of the new contingency is not included in 1-2-3. The fourth (2-4-6) contingency was designed to be the hardest, since the animals would have to skip an arm to get to the correct outer arm of the contingency. The fifth (2-3-4) and sixth (4-5-6) contingencies were chosen for comparison with the first three contingencies to understand the evolution of the ability of the animals to perform the task and generalize from previous experience. As opposed to behaviors designed to study asymptotic performance, there need not be strict criteria for switching between the contingencies since the purpose of this task was to understand the continual learning and behavior of the rats. Furthermore, the automated systems matched the number of inbound rewards of the animals, ensuring that all animals had similar learning opportunities. We therefore switched to a new contingency the day after >80% of the animals received >80% reward over the course of a session. That ensured that by the time each contingency switched almost all of the rats reached at least ~80% correct on a session during each contingency (Fig S2C).

Performance improved on each of the contingencies, such that by the end of each one, almost all rats made few outbound or inbound errors (Fig 1G, S2C). There was, however, substantial systematic variability across animals. For instance, we found that rats that behaved poorly in a single contingency also behaved poorly in other contingencies, and vice versa (see colored lines in Fig 1F&G). This variability provided an additional goal for our modeling, in that an ideal model would capture not only the overall learning of the group but also the differences among individuals. Importantly, just because the rats show differences in their behavior does not mean that those differences are due to innate capabilities of the rats. The variability could occur due to a form of path dependency—enshrining aspects of the idiosyncratic vagaries of

their initial engagement with the task in their later behavior. Our modelling approach has at least some capacity to separate these innate and experiential causes of variability.

Model Fitting

The procedures of model generation, testing and refinement are exactly those of formulating and falsifying hypotheses about possible underlying causes of the observed behavior. To put it another way, the hypotheses are typically componential—involving different potential processes (such as learning rules) and parameters (such as subject-specific learning rates or weights governing the impact of alternative mechanisms). It is these that are captured in the models. Since these are often somewhat buried or ignored in modeling papers, we take the opportunity to lay out the logic of their development and motivate their form through the intricacies of the data. Testing such hypotheses requires determining the values of the parameters that fit the data, and balancing the quality of the fits against the complexity of the models (Daw, 2011). New and additional hypotheses then arise from features in the residuals of the fits to the data. We illustrate these procedures by presenting a series of related models that increasingly capture the behavior of individual animals across multiple stages of the task. Since, by their very nature, models offer incomplete and simplified representations of phenomena of interest, we then address how and where to stop the formal modeling process, and what we learn from the final, albeit inevitably partial, model.

As a choice of modeling framework, we sought a simple algorithm that did not require acausal information, would alter its internal information based upon its choices and rewards to increase the expected return of reward, and could work in the face of partial observability (see below). Therefore, we used the actor-critic class of RL accounts trained by the REINFORCE policy gradient algorithm (Williams, 1992). This class is a popular choice for characterizing animal learning behavior in RL paradigms (Suri and Schultz, 1999), and there is also evidence of the use of REINFORCE-like methods in humans (Li and Daw, 2011).

According to these accounts, an animal's choice of an arm on trial t, which we write as a_t, depends probabilistically on an internal characterization of its situation or state s_t, which can contain various sorts of information such as past arm choices. This dependence arises through a collection of action preferences or propensities $m(a, s_t)$, such that actions with higher propensities are more likely to be chosen. The propensities are updated in the light of rewards. The full details of the equations involved are provided in the Methods. However, in brief: a conventional softmax function converts the propensities to probabilities, $p(a; s_t)$, of choosing to go to arm $a_{t+1} = a$ on this trial (Eq. 1). Via the rules of the task, this choice of arm then determines whether the model receives a reward, r_{t+1}, and also causes the state to update to s_{t+1}. This reward is then used to calculate the prediction error, δ_t, using the value function of the critic at states s_t and s_{t+1}, $V(s_t)$ and $V(s_{t+1})$ (Eq. 2). δ_t is then used to update $V(s_t)$ (Eq. 6) and the factors governing the propensities $m(a, s_t)$ (Eq. 3 – 5). Finally, new propensities $m(a, s_{t+1})$ are calculated, at which point the process begins again (Fig 2A).

For the fit to an individual rat, we forced the model to make the same sequence of arm visits as the animal during the initial exploratory phase, effectively using the data of the animal to inform the initial condition of the model. Additionally, we chose to fit specifically the second and third contingency, and then to see how those parameters caused the model to behave on

the rest of the contingencies. The second and third contingencies are the most representative for this task, as 1) they both follow other simple contingencies, and 2) occur before the hardest, fourth, contingency, for which the required alternation involves skipping neighboring arms. Furthermore, fitting the second and third contingency allowed us to use the performance of the model on the subsequent, and preceding, contingencies as predictions that could test the goodness of fit of the models.

A Basic RL model cannot learn spatial alternation task

In the first model (M1), we chose a simple formulation that relates choice to current location. In M1 the state is defined by the current location, $s_t = \{a_t\}$, such that $m_1(a, s_t) = b(a|a_t)$. For each state, i.e. current arm location, $b(a|a_t)$ contains 5 numbers governing the propensity to make a transition from the current arm to each of the other 5 arms. Returning to the same arm is not allowed in the model, as it was never rewarded in the behavior. The model has only three parameters, all of which take values between 0 and 1. The first parameter is the temporal discount factor, γ, which determines the weighting of rewards in the farther future in defining the long-run values of states (and thus in calculating the prediction error, δ) (Eq. 2). The second parameter is the learning rate, α, which determines how much δ updates the propensities and the value function (Eq. 3 – 6). The third parameter is the forgetting rate, ω, which determines how quickly the propensity parameters and the value function decay towards 0 (Eq. 3 – 6), a value that would indicate that there is no specific information about which arm to visit in any state. ω enables the model to learn the nonstationary task by constantly depreciating old information, allowing for more rapid changes in the propensities and value function during changes in contingencies.

To evaluate the theoretical capability of M1 to perform the same task as the rats, we found the set of parameters that maximized the amount of reward that the model could receive after initializing the model with the sequence of exploratory behavior that the animals take. That M1 is inadequate at performing the task can be seen through comparing its performance to the average behavior across all of the rats (Fig 2B). Although M1 improves over the course of each contingency, its maximal reward at any given contingency does not reach above ~65%, and so remains well below the performance of the animals.

In order to fit the model to characterize the behavior of each individual animal, we sought a procedure that could find parameter values that minimize the discrepancy between the choices the model would make and the actual choices of that animal. As is conventional, we first found the parameters that maximized the likelihood that the model chose the exact arms that the animal selected. However, when the model, with these parameter values, is allowed to generate new choices by itself (as a form of generative test (Wilson and Collins, 2019)), simple statistics of the inbound and outbound errors failed to match those of the animals (Fig 2C). By comparison, generating new choices based on the parameters that maximized the amount of reward that this model could get on the task did far better at recapitulating these error statistics. This was not an error in our fitting procedure since if we generate arm visits from the set of parameters that maximized the amount of rewards the model could receive; we were able to recover those same parameters using just this ML method (Fig 2C&D). Instead, the

deviation from the ML fit is a product of the inability of the model to fit the exact choices of the rats sufficiently accurately.

That ML fails in the face of model misfit is not surprising; however, its inability to find parameters that reproduce obvious characteristics of the behavior make it hard to use to investigate the structure in the residual error of the model, as it is this structure that points the way to improve the model. Therefore, instead of the ML method for fitting, we used an Approximate Bayesian Computation (ABC) method (Lintusaari et al., 2017) which finds parameters such that the average behavior of the model when operating in the task, choosing stochastically, matches as well as possible that of an individual animal, according to some suitably-chosen statistics. We averaged over 200 repeats of the model and chose as statistics the inbound and outbound performance for the second and third contingencies. The ABC parameters were found to minimize the root mean squared (rms) difference between the model and rat inbound and outbound performance on these contingencies. Compared with ML, the ABC method when fitting to the behavior of individual animals did a better job of capturing the errors rates of the animals, and, perhaps not surprisingly, found sets of parameters far closer to those that maximized the reward that M1 could gain (Fig 2C&D). As we will show below, the comparison of the parameters from the ABC fits to the parameters that maximize the reward of a given model provides a helpful rubric for understanding if a given model can capture innate variability, as opposed to random variation, in performance.

As expected, even with this alternative fitting method, M1 cannot perform this task well: it treats the reward contingency as being Markov, depending only on the current arm (Sutton and Barto, 1998). However, at the center arm of a contingency, just knowing the current arm cannot uniquely define the next correct arm to visit, as that requires knowledge of the previous arm visited. In formal terms, the reward contingency is partially observable (formally, a partially observable Markov decision process or POMDP; (Kaelbling et al., 1998)) or aliased—the immediate sensory information (the current arm) does not suffice to inform appropriate behavior. This is why the model does fairly well on capturing the inbound trials of the individual rat, but utterly fails at capturing the outbound behavior.

Working memory RL model learns task slowly

One way to solve a POMDP is to enrich the notion of state s with historical information that resolves aliasing, effectively generating an augmented state which does observes Markovian dynamics. For model M2, we did this by adapting and simplifying for our task an actor described by Todd et al. (2009). This adds to the state a memory unit that can store past actions. This enables M2 to make decisions based upon current and past information. Such a strategy had been used to learn common rat behavioral tasks (Zilli and Hasselmo, 2008), and exhibits features of rat behavior (Lloyd et al., 2012). In M2, the state, $s_t = \{a_t, a_{t-1}\}$, includes both the current and the most recent past arm (Fig 3A). For this model the propensities are now $m_2(a, s_t) = b(a|a_t, a_{t-1})$. In all other respects M2 is identical to M1.

This working memory (WM) RL agent is now able to learn the task; however, in large part, it does so far slower than the rats, even for the parameters that maximize the total reward (Fig 3B). With M2, good performance on the first contingency arises at the correct timescale—something that will be discussed further below—but performance on all the subsequent

contingencies improves much more slowly than for the rats (Fig 3B&C), including the second and third contingencies on which it was fit to the behavior. As for M1, the parameters from fitting all animals to M2 are close to the parameters that maximize the amount of reward that the agent can receive, and thus do not capture individual differences due to innate capabilities among the animals (Fig 3D).

It is worth mentioning, that even though this model contains the capacity to maintain past information—i.e. has memory—it cannot learn the task as rapidly as the animals. These spatial alternation tasks are often used as a means to probe memory and changes in memory with different manipulations; however, the failure of this model to capture the rapidity of the learning of the rats highlights that other components could be responsible for the learning of the task, ones that are less closely associated with working memory.

Adding arm and transition preferences enables the model to learn as rapidly as the rats

We then considered additional information that could help the model perform more similarly to the animals. As shown above, the animals exhibit biases that are unrelated to reward contingencies in that they do not visit the different arms randomly during the exploratory phase of the behavior (Fig 1B&E). We therefore asked if such dynamic biases could underpin their later swift learning.

To capture these biases we added terms to the propensities such that for model M3: $m_3(a, s_t) = b(a|a_t, a_{t-1}) + b^i(a)$. The term $b^i(a)$, which we call a dynamic independent arm preference, provides the agent with additional preferences to choose specific arms next, independent of its current or past locations. As with the state-dependent propensity terms, $b^i(a)$ are also updated by δ_t through the process of learning (Fig 4A). Importantly, adding this term allows us to capture both the fact that the animals may prefer specific arms before they begin the learning of the alternation contingencies and that these preferences can be shaped by reward

Adding the dynamic independent arm preference makes M3 drastically faster than M1 and M2, coming far closer to the performance of the rats (Fig 4B&C). However, as also seen with M2, the first contingency still diverges from the real behavior, now with M3 learning faster than the animals. For M3, the values of α and γ, but not ω, for all of the individual animals are close to the values that maximize the amount of reward that M3 can receive (Fig 4D), indicating that M3 still does not help us determine whether the differences among the behavior of the individual rats is due to differences in innate capabilities.

Finally, we added two dynamic transition preferences, creating model M4 (Fig 5A), for which $m_4(a, s_t) = b(a|a_t, a_{t-1}) + b^i(a) + b^{n_1}\chi_1(a = a_t \pm 1) + b^{n_2}\chi_2(a = a_t \pm 2)$. The additional propensities capture the preference of the animals seen in figure 1C to transition to neighbors that are either one, b^{n_1}, or two arms b^{n_2}, away. Again, these are both also updated by δ_t. It is worth noting, that although we have increased the internal complexity of the model, it still only has the same three parameters as the initial model. All of the factors in the propensities are updated with the same learning, α, and forgetting, ω, rates.

With the addition of these dynamic independent arm preferences, and dynamic transition preferences (Fig S3), model M4 fits contingencies two and three well (Fig 5B&C). In

addition, with M4 there was substantial variability in the parameter values across animals, which were no longer all the same as those that maximize the amount of rewards that the agent can receive (Fig 5D). This suggest that M4 could capture and expose individual differences due to innate learning. We expand on this further below.

To quantify the improvement in fit of M4 over the other models, we calculated the root mean square difference between the model and data for all rats for all inbound and outbound trials during each contingency (Fig 5E). We found that adding the dynamic independent arm preference dramatically improved the rms difference in both the inbound and outbound trials. This makes sense as the arm preference dynamically learns the three arms that provide reward in a given contingency (Fig S3A), effectively decreasing the space of states that the model has to explore to learn the contingency. Adding the dynamic transition preferences further improves the fit to the outbound trials for the same reason, since, now, when the model is at the center arm, it is far more likely that it will visit a neighboring arm.

Generalization across contingencies

The parameters of the models were chosen to fit the data from the animals on just the second and third contingencies. To quantify how well M4 generalized to the remaining contingencies we first computed the rms differences between the inbound and outbound error statistics of single runs of the model and the average behavior of the model (Fig 6A). These provide a baseline for interpreting the rms difference between the model and the animals, as they allow a comparison, for each contingency, of the variability in error statistics inherent in the model and the differences between the model and the behavior of the individual rats. Fig 6B shows this measure for the outbound errors for the fourth contingency. This allowed us to define the fit accuracy as the probability of the simulations of the model having a larger error than the data and to examine the distribution of fit accuracies across contingencies (Fig 6C).

Examining these distributions revealed clear successes of the model. The first, and possibly most surprising, is its ability to predict the outbound errors of the fourth contingency. This contingency, 2-4-6, is by far the hardest contingency, as it forces the animals to continue to alternate arms but to do so whilst skipping neighboring arms. The ability of the agent to capture such a distinct contingency suggests that the dynamic transition preferences capture key aspects of the animals' behavior.

Importantly, this fit was by no means guaranteed. One could imagine that in terms of these outbound errors on the fourth contingency, the rats could have performed either far better or worse than the model. If the rats had generalized the fact that they had to alternate, then they would have performed far better than the model, which has no capacity to generalize the structure of the task. Alternatively, the rats could have developed a model whereby they only get rewarded at arm 4 but have to visit another arm prior to getting that reward, in which case the model would have done far better than the rats.

The second success is that the agent predicts the inbound errors for the fifth and sixth contingencies. Both of these contingencies are of the more standard variety, with neighboring arms being used in the alternation. As with the above example, this suggests that the model has

extracted core aspects of the way in which the rats transition between contingencies and continually learn.

The success is again informative because the animals could have performed better or worse than the model. The animals would have done better if they had learned that they live in a world that changes contingencies and were able to detect when that was the case more rapidly and adjust their behavior accordingly. M4 lacks the capacity to realize or act upon such regularities. The rats could also have performed worse than the model if they had been confused by the return from the skip arm contingency and had to relearn something to continue to perform the task well.

The failures of the model are equally, if not more, informative. The first failure of the model is apparent at a time when the animals' knowledge might have been destabilized by the structure of the task. This is during the inbound trials in the fourth contingency. In the transition from the third contingency, 3-4-5, to the fourth contingency, 2-4-6, the center arm does not change, which is why the model makes essentially no inbound errors at that transition. The animals, on the other hand, systematically show an increase in inbound errors after the transition to the fourth contingency. One could imagine many ways that this could happen. For instance, this could be an indication of a global destabilization of information whereby, since the animals have a difficult time determining how to get rewards, they discard other things that they already know, even though they might be correct. Equally, it could be a signature of a form of directed or undirected exploration that we do not model.

Second, the distribution of fit accuracies shifts for the worse for the outbound errors of the fifth and sixth contingencies. The median of the distribution is 0.092 and 0.099, respectively (Fig 6C). These medians are far worse than the outbound fit accuracies for the second (0.363) and third (0.752) contingencies, and, surprisingly, they are also worse than the outbound fit accuracy for the fourth contingency (0.237), even though the fifth and sixth contingencies are of the standard variety with neighboring arms being the outer arm.

To better understand the nature of this shift in the fit accuracy for the outbound errors of the fifth and sixth contingencies we calculated the average difference between the errors of the individual rats and the errors of the model fits (Fig 6D). These residuals confirm the observations from the distributions of the fit accuracies. The model makes fewer errors, inbound and outbound, on the first contingency, and it also makes fewer inbound errors on the fourth contingency. However, the model makes more outbound errors at the beginnings of the fifth and sixth contingencies. We view this difference as a potential indication that the animals have generalized something about the structure of the task at this point, such that they know that they have to alternate, they just have to find the center arm from which they will then alternate, leading to effectively no outbound errors upon switching contingencies.

Finally, and as we have seen with the various models, the first contingency is quite different. We will return to this first contingency and the differences between the model and the rats below.

Parameters of individual model fits capture variability in behavior

With model M4 and with the addition of the neighbor arm preferences, the parameters of the fits to the individuals showed more variability and deviation from the parameters that maximized the reward (Fig 5D). When compared to M3 the M4 fits to all 24 rats had an interquartile range 3.0 times larger for γ (0.11 vs. 0.04), 7.8 times larger for α (0.39 vs. 0.05), and 1.8 times larger for ω (0.004 vs. 0.002). Given the presence of this variability, we asked whether it could reflect the variability in performance that we see in the data (Fig 1F&G).

First, we need to ensure that the differences in the parameter values we get from the fits are meaningful. We therefore evaluated the error landscape of the fits to determine whether there were clear global minima for each animal. We found that there were indeed global minima that were distributed across the parameter space. Our fitting procedure reliably determined the vicinity of the global minima (see Fig S3A for an example), indicating that the differences among animals are meaningful.

These different global minima also corresponded to different reward rates. We computed the average reward rate across 200 repeats of the model for a large range of parameters and found that the reward rate peaked at a large learning rate, small temporal discounting and midrange forgetting rate. We then compared that reward space to the range of the parameters of multiple fits to the individual animals (Fig 7A). The variability in the parameters corresponded to different reward rates for the model, indicating that the different model fits could reflect different learning capabilities in the rats.

To represent the variability in the behavior of the animals, and to see if the model captures that variability, we calculated the reward rate for all animals across the second and third contingencies, those contingencies fit by the model. As a comparison we also calculated the average reward rate of 200 repeats of the model for each of the individual fits of the model also for the second and third contingencies (Fig 7B). Both reward rates varied across a large range, and those animals with the lowest behavioral reward rates also had the lowest model reward rates. However, the model and behavioral reward rates differ in two ways. The first is that the highest performing rats were able to receive more rewards than could the model. This exposes another limitation of the model. Given fixed parameters for the entirety of the behavior, a good fit of the model has to find the right balance between, for example, a fast forgetting rate necessary for rapidly learning the different contingencies, but not too fast that the model forgets information during the performance of a single contingency and gets less reward at its asymptotic behavior. A possible way forward would be to incorporate dynamic parameters.

The second way in which the behavioral and model reward rates differ is that the ordering of the animals with increasing reward rate is not identical between the behavior and model reward rates. This could be viewed as another point of failure for the model; however, it is also possible that the model has extracted a more abstract version of the learning capacity of the animal; if so, this difference in ordering is a success of the modeling. To distinguish these two possibilities, we performed a median split of the animals based upon their actual reward rates (Fig 7C) and their model reward rates (Fig 7D) over the second and third contingencies and averaged the actual behavior of the two groups across all the contingencies.

If the models have abstracted information related to each animal's learning capacity, then this median split should better separate groups on contingencies 1, 4, 5, & 6 (those not taken into account for the median split). That was indeed the case: the average difference between the two groups (Fig 7E) when split along the median using the model reward rate was 40% larger than when split along the median using the actual reward rate (7.5 % vs. 5.4%) ($p = 0.008$; permutation test). Therefore, the model reward rate more veritably reflects the overall learning capacity of the rats.

Exploratory behavior persists into the first alternation contingency

We now return to the first alternation contingency, which has stood out at multiple stages in the modeling process. One explanation as to why that might be the case comes from considering M4 and the steps taken to develop the model to that point (Fig 5).

Model M4 captures the speed of the learning of the subsequent contingencies but learns the first alternation contingency too quickly. It is possible that we have not added sufficient preferences and biases to the model. To create M4 we added preferences that would enable the model to speed up its learning; however, the rats also show directional inertia (Fig 1D&E), and that could work against their ability to learn the task rapidly. This raises the hypothesis that the rats show directional inertia that persists into the first alternation contingency. This hypothesis is consistent with the entirety of this task drawing upon previously unappreciated aspects of the behavior of the rats beyond just memory storage.

To understand what else could be missing in the model, and thereby what other aspects of the behavior of the rats might govern their learning, we reconsidered the way in which the rats interact with the arms during the exploratory period (Fig 1) and examined the neighbor frequency and directional inertia (Fig 1 C – E) of the animals and model during the exploratory period and into the first alternation contingency. We compared the behavior of the rats to the average behavior of M2 and M4, using the parameters from the fits to the individual animals. M2 provided a useful foil, as that model did a better job capturing the rate of learning of the animals during the first alternation contingency.

We calculated the frequency of transition to a neighboring arm in each session (Fig 8A). The values are identical between the animals and the models during the exploratory period because we force each model to follow the same series of arm visits as the individual rats (see Methods). Following the transition to the first contingency, M2 and M4 show distinct behavior. M2 shows a more precipitous decline and a slower return to asymptote in the neighbor frequency when compared to M4. The rats behave in a more consistent way with M4. Their initial decrease is larger than that of M4, but they do not persist at the low level, rapidly increasing to their asymptotic performance.

The fact that the rats show more consistent behavior with M4 supports the hypothesis that even the first contingency is not exclusively a memory-based task since in the purely memory-based version of the model (M2) this preference for neighboring arms develops much slower than in the animals (and in M4). However, that does not explain why the animals are so much slower to learn this first contingency than M4. To understand what might be at play we turned to the directional inertia of the animals.

For the directional inertia (Fig 8B) M2 and M4 show relatively similar dynamics, both of which are distinct from the rats. At the transition to the first contingency, both of the models initially drop and then rise to their asymptotic directional inertia levels. In contrast, the rats persist with an elevated directional inertia after the transition to the first alternation contingency, and then drop and eventually rise back up to their asymptotic directional inertia level. Neither M2 nor M4 has any additional preferences to capture the directional inertia, and this persistent value in the rats can indicate that, at least part of the slowness of the animals, can be due to the persistence of this pre-existing bias.

These results indicate that that persistent directional inertia could be a cause of the slower learning of the first contingency. To evaluate that possibility, we split the neighbor frequency and directional inertia metrics between the high performing and low performing animals as dictated by the median split of the model reward rate (Fig 7D). If directional inertia is related to learning rate, then the low performing animals should show more persistent directional inertia than the high performing animals (Fig 7D). That, indeed, is the case. The lower performing animals show an increased persistence ($p = 0.004$) of the elevated directional inertia (Fig 8D) into the first alternation contingency, but do not show different numbers of trials to rise in their return towards asymptote ($p = 0.08$) for the neighbor transition frequency (Fig 8C).

One of the intriguing hypotheses raised by exploring the variability between rats (Fig 7) is that the rats could have different innate abilities to learn. Those rats that learn the second and third contingency the best, as defined by the model, are also the rats that learn most of the other contingencies the best. This same grouping shows differences in the neighbor transition frequency during the exploratory period ($p = 0.006$, permutation test) (Fig 8C). The high performing rats show a greater neighbor transition frequency during this period. It is interesting that it is the neighbor transition frequency that differs in the exploratory period, as that was the feature that we had to add to the model in the first place to enable it to encompass some of the variability amongst the rats (Fig 5D).

Not only does the neighbor frequency bias differ between the groups of high and low performing animals, but the average neighbor frequency during the exploratory period of the individual rats shows a very high correlation ($r^2 = 0.63$, $p = 5 \times 10^{-6}$) with the amount of reward that same rat receives during the first alternation contingency (Fig 8E). This further supports that the pre-existing biases and dynamic preferences of the rats are a major contributing factor to their behavior, even during the first alternation contingency.

It is important to note that this connection between the exploratory behavior and the learning during the first spatial alternation contingency is not a causal statement. Further experiments will be necessary to understand if the higher performing animals perform better on the first alternation contingency (and subsequently contingencies) because they have a larger neighbor bias during exploration or if there is something about a larger neighbor bias that relates to greater learning capabilities in general. For instance, when innate learning capabilities have been studied in mice, there has also been a connection between faster learners and exploratory behavior (Matzel et al., 2006).

Discussion

We have presented a behavioral and modeling paradigm to advance our understanding of the nature of continual learning in rats. Using an automated behavioral system (Fig 1A), we recorded the continuous behavior of multiple rats. The rats learned a series of spatial alternation contingencies, whereby the next rewarded arm could be a function of not just the current location of the animal, but also on the past behavior of the animal (Fig 1F&G). Instead of beginning the learning of the alternation task with the initial exposure of the animals to the track, we separated the learning of the environment from the learning of the task by starting all animals on an exploratory period of the behavior. The exploratory period demonstrated that the rats interacted with the apparatus by showing multiple components of non-random behavior, exposing their pre-existing biases and the substrate for dynamic preferences (Fig. 1B – E).

Through refinement, we developed an RL agent that could perform the task as quickly as the rats and fit that model to capture the individual behavior. We laid bare the logic of the changes that we made to the RL agent to develop the final model. As the alternation task cannot be described as a Markov decision process, the simple model-free RL agent (M1), with which we started, could not learn the task (Fig 2). A working memory RL agent (M2) (Todd et al., 2009) could learn the task, but did so far slower than the animals for all but the first alternation (Fig 3). Incorporating dynamic preferences, motivated by the exploratory behavior of the rats, into the RL agent (ultimately model M4) enabled it to learn as rapidly as the rats (Fig 4&5). Note that M2 and M4 can both be described as model-free agents; thus our result shows that one need not necessarily appeal to more computationally sophisticated, model-based, components (Gershman and Niv, 2010) to account for all aspects of fast learning, but rather take appropriate account of structural contributions associated with biases.

Through fitting M4 to a subset of the alternation contingencies we compared the performance of the agent to the individual rats. The RL agent captured some aspects of the learning of the rats well, with the biases underlying M4 explaining different aspects of the individual variability of the animals (Fig 8). Perhaps the most surprising match between M4 and the rats was its ability to predict the outbound errors during the hardest alternation contingency (Fig 6). However, M4 did not perfectly capture all aspects of the behavior. There were systematic differences between the model and rat behavior. Of perhaps most interest is that the model made more outbound errors on the final alternation contingencies (Fig 5&6). We speculate that it is here that M4 is compromised by its inherent model-free nature and that the rats outperform the RL agent at these later contingencies because by that time they have generalized that they are performing a task that alternate and could use that information to speed up their learning of a given contingency. It will be interesting to study the timing and nature of this potential generalization further. Does the generalization only happen due to the number of contingencies, or is there something about the animals' experience of the skip arm contingency that allows them to generalize the alternating nature of the task more competently?

For this behavior we chose a simplified output as the modeled feature: visiting arms. The nature of the algorithm that governs the behavior led to the choice of arm visits for the model, as arms visits are the only factor taken into account when evaluating rewards. However, there are clearly many additional aspects of the behavior of the rats that are not completely

subsumed by the arm the animal visits. Future models will be necessary to account for the timing of arm visits, the precise trajectory the animal takes, and whether or not the animal breaks the beam of the reward well, as well as other potentially relevant ways in which the rats interact with the space of the task.

Modeling considerations: ML vs ABC and when to stop modeling

In developing and fitting the RL agents we had to make many decisions, many of which highlight general points about modeling animal behavior. When fitting individual animals with the model we found that maximizing the likelihood that the model would chose the same arms as the animal did not provide parameters that would then be able to generate statistics that captured the way in which the rats behaved (Fig 1C – D). Therefore, consistent with other studies using RL agents to fit rodent behavior (Lloyd et al., 2012; Luksys et al., 2009), we utilized ABC methods to minimize the difference between characteristic statistics of the rats and the model. This is not formally surprising, since ML offers no guarantee for models that underfit data; but it would be important to understand such failures more generally, since by far the bulk of behavioral modeling employs methods that are vulnerable to the same concern.

Through the various model iterations, we had to decide both when to continue refining the model, and when to stop doing to so. The initial drive to continue the modeling was obvious. M1 couldn't perform the task (Fig 2) and M2 couldn't learn the task as rapidly as the animals (Fig 3). However, with M3, the need to continue was far more subtle. M3, largely, had the capacity to learn the task as rapidly as the rats; however, when we fit M3 to the individual animals the parameters mostly clustered around the parameters that maximized the amount of reward that M3 could receive (Fig 4D). This indicated that there was still something fundamentally missing in the agent. None of the subtlety and richness of the individual variability between animals existed within the parameters for the fits of M3. Finally, with M4, the parameters from the fits varied away from the maximal parameters from that agent. This indicated that we could consider no longer continuing the process of improving the modeling.

The decision to stop the modeling at that point rested upon somewhat different factors. Even though M4 fails to capture certain identifiable aspects of the behavior of the rats, it, surprisingly at times, closely matches different features of the way in which the individual animals learned the task (Fig 6), and better separated them than simple statistics such as the reward rate (Fig 7). It was the combination of capturing variability across animals and predicting features of the behavior outside of the fit that suggested that we might stop refining the model further. To go beyond this, it would be desirable to refine the paradigm, for instance to put the 'skip' contingency (2-4-6) at different points to examine its role in generalization; or to systematize the length of engagement with each contingency to study the possible decrease in learning rate as performance improved.

Modeling functions: quantitative understanding for hypothesis generation

The success of Model M4 highlights a number of factors about the behavior of the animals. First, it sets something like a lower bound on the information and strategy an animal could use to perform the task. M4 sets into relief the fact that the rats, for much of their behavior, need not know anything more complicated than a working-memory-based variant of

a model-free RL agent. That, of course, does not mean that they do not know anything more, it just shows that they need not to perform the behavior the way that they do. M4 also highlights that this spatial alternation task is not just a memory task. Rather, they are, at least as much, a study in the way in which rats utilizes and alter their behavioral biases.

Second, in the places M4 fails to capture the behavior it can highlight the aspects of the task that are least well understood to direct further inquiry. When this lower bound cannot keep up with the behavior of the rats, as evident in M4's characterization of the outbound errors of the fifth and sixth contingencies, it focuses us on timepoints worthy of further experimentation.

Third, animals behave differently, something lost in averaging behavior across animals. This individual variability can be due to many different causes, but at the extremes are that random variation underlies that variability between animals or that the variability is due to innate differences between animals. These two extremes are far from mutually exclusive and very difficult to tease apart, especially with a continual learning task. Individual rodents show consistent learning ability across tasks (Galsworthy et al., 2002; Matzel et al., 2003), indicative of there being innate differences between animals. M4 was able to capture variability between animals (Fig 7), and it served as a better way to separate out that variability than the actual reward rate of the rats, suggesting a reality to what it captured.

Conclusion

In sum, we have shown the promise and problems of modeling relatively complex behavioral patterns in rodents. We examined the course of learning from the very outset in some detail, and thereby gained insight into aspects of the different strategies employed by our subjects. These results point directly to further questions that new variants of the paradigm can duly elucidate.

Methods

Animals: All experiments were conducted in accordance with University of California San Francisco Institutional Animal Care and Use Committee and US National Institutes of Health guidelines. Rat datasets were collected from Long Evans rats, ordered from Charles River Laboratories, that were fed standard rat chow (LabDiet 5001). To motivate the rats to perform the task, reward was sweetened evaporated milk, and the rats were food restricted to ~85% of their basal body weight.

Two cohorts of rats, comprised of 6 males and 6 females each, were run on the automated behavior system. There were no systematic differences in reward probabilities between the male and female rats within the two cohorts (Fig S2A), so data from all animals were aggregated for subsequent analyses. The entire behavior took place over the course of 22 days for the first cohort and 21 days for the second cohort. The first cohort ran an extra day on the initial exploratory behavior, where the animals received rewards after visiting any arm of the track. At the start of the behavior the first cohort of rats were 4 – 5 months old, and the second cohort of rats were 3 – 4 months old.

Automated behavioral system: The automated behavior system was custom designed and constructed out of acrylic. All parts of the behavior system were enclosed with walls. There were different symbols on each arm of the track serving as proximal cues, and there were distal cues distinguishing the different walls of the room. Pneumatic pistons (Clippard) opened and closed the doors. Python scripts, run through Trodes (Spike Gadgets), controlled the logic of the automated system. The reward wells contained an infrared beam adjacent to the reward spigot. The automated system used the breakage of that infrared beam to progress through the logic of the behavior (Fig S1A). In addition to the infrared beam and the spigot to deliver the reward, each reward well had an associated white light LED (Fig S1A).

Each cohort of rats were divided into groups of four animals. The same groups were maintained throughout the duration of the experiment. Within a group, a given rat was always placed in the same rest box, and the four rats of a group serially performed the behavior. The rats had multiple sessions on the track each day. During the exploratory period of the behavior, the duration of a session was defined by a fixed number of rewards. During the alternation task the duration of a session was defined either by a fixed number of center arm visits and at least one subsequent visit to any other arm, or a fixed amount of time on the track (30 minutes), whichever came first. Only one of the female rats reached the time limit, and it did so for only two sessions toward the beginning of the first alternation contingency. For that one female we incorporated the trials that she ran on those sessions and did not distinguish the time out sessions for the analyses.

The algorithm underlying the exploratory part of the behavior had only one rule. Reward was delivered for any infrared well beam break if and only if the current well infrared beam break was immediately preceded by an infrared beam break at any other well. This prevented the animals from getting continuous reward at a single arm, and ensured the rats visited at least two of the arms.

The algorithm underlying the spatial alternation task was such that three arms on the track had the potential for reward within a given contingency, for example during contingency at arms 2-3-4, arms 2, 3, and 4 had the potential to be rewarded, and arms 1, 5, and 6 did not. Of those three arms we will refer to the middle of the three arms as the center arm (arm 3 in the above example) and the other two arms as the outer arms (arms 2 and 4 in the above example). Reward was delivered at the center arms if and only if: 1) the immediately preceding arm whose reward well infrared beam was broken was not the center arm. Reward was delivered at the outer two arms if and only if: 1) the immediately preceding arm whose reward well infrared beam was broken was the center arm, and 2) prior to breaking the infrared beam at the center arm, the most recently broken outer arm infrared beam was not the currently broken outer arm infrared beam. The one exception to the outer arm rules was at the beginning of a session, following the first infrared beam break at the center arm, where only the first condition had to be met.

For the running of the behavior, the infrared beam break determined an arm visit (Fig S1A); however, the rats would sometimes go down an arm, get very close to the reward wells, but not break the infrared beam. Therefore, for all of the analyses described, an arm visit was defined as when a rat got close to a reward well. These times were extracted from a video

recording of the behavior. These, effective missed pokes were more frequent at the beginning of a contingency (Fig S2B), but overall were not that common. This proximity-based definition of an arm visit added additional arm visits to those defined by the infrared beam breaks, and by definition none of them could ever be rewarded, nor alter the logic of the underlying algorithm. However, because of the non-Markovian nature of the reward contingency, they could affect the rewards provided for subsequent choices.

RL agents: Given that each spatial alternation task could be framed as a partially observable Markov decision process, we adapted the working memory model of Todd et al. (2009) as the basis for our series of RL agents. The models specify rules governing propensities $m(a, s)$ that contain the preferences of the agent of choosing arm a when the state is s. Models differ in terms of what counts as the state, and also according to the various terms whose weighted sum defines the propensity.

In the first agent (Fig 2) the state is defined as the current arm location, $s_t = a_t$, of the agent. In all subsequent agents the state is defined as the combination of the current arm location of the agent and the immediately preceding arm location of the agent, $s_t = \{a_{t-1,} a_t\}$. This is a simplification from the Todd et al. model, whereby a_{t-1} is always placed into the memory unit, effectively setting the gating parameter for the memory unit to always update the memory unit. Then, the first component of $m(a, s)$ for all models is $b(a, s)$, which is a 6x(6+1) or 6x(36+6+1) matrix containing the transition contingencies to arm a from state s. The reason for the additional states beyond just the 6 arms or 6x6 arms by previous arms is to include the rest box in the possible locations to allow for the inclusion of the first arm visit of a session. In so doing that adds 1 additional state to model M1, and 6+1 additional states into the subsequent agents since the animals can be located in the rest box and can be located at any of the 6 arms having previously been in the rest box.

To provide the agents with additional spatial and transitional preferences we added components to the transition propensities. The first is an arm preference, $b^{\mathrm{i}}(a)$ that is independent of the current state of the animal. The second is a preference for visiting arms that neighbor in space the current arm, $b^{n_1}\chi(a = a_t \pm 1)$, where $\chi()$ is the characteristic function that takes the value 1 if its argument is true (and ignoring arms outside the range 1…6) and b^{n_1} is the (plastic) weight for this component. The third is a preference for visiting arms that are two removed, in space, from the current arm, $b^{n_2}\chi(a = a_t \pm 2)$. The neighbor arm preferences contain only single values, the preference to go to a neighboring arm. These neighbor biases were applied equally in both directions when possible (i.e. if the agent was at the end of the track the neighbor bias could only be applied to one direction).

To determine the probability of visiting each of the arms from a given state, the total propensity is passed through a softmax such that:

$$p(a; s) = \frac{\exp(m(a, s))}{\sum_b \exp(m(b, s))} \tag{1}$$

The agent's visit is the determined by a sample from this distribution. The choice of arm then determines the reward, r, which is either 0 or 1, based on the algorithm that governs the

spatial alternation task. The probability of revisiting the current arm is set to zero, and the probabilities of going to the remaining arms sums to 1.

The model uses the REINFORCE policy gradient method (Williams, 1992) within the actor-critic framework of temporal difference learning, to update the propensities in the light of the presence or absence of reward. To do this, the agent maintains a state-long-run-value approximation, $V(s)$, which functions as a lookup table, with one component for each state. The reward determines the state-value prediction error:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{2}$$

where $\gamma \epsilon [0,1)$ is a parameter of the model called the temporal discounting factor, which determines the contribution of future rewards to the current state.

δ_t is then used to update the preferences all of the components of the propensities and $V(s)$. The state-based transition component is updated according to the rule:

$$b(a,s) \Leftarrow b(a,s)(1-\omega) + \alpha\delta_t \times \begin{cases} 1 - p(a;s), s = s_t, a = a_t \\ -p(a;s), s = s_t, a \neq a_t \\ 0, s \neq s_t \end{cases} \tag{3}$$

where $\alpha \epsilon [0,1]$ is a parameter of the model called the learning rate, which determines the amount by which all components of the propensities change based on the new information. $\omega \epsilon [0.001, 0.015]$ is also a parameter of the model called the forgetting rate, and determines how the propensities associated with unchosen states (and/or actions) decay. The independent arm preference is updated according to the rule:

$$b^{\mathrm{i}}(a) \Leftarrow b^{\mathrm{i}}(a)\,(1-\omega) + \alpha\delta_t \times \begin{cases} 1 - p(a;s), a = a_t \\ -p(a;s), a \neq a_t \end{cases} \tag{4}$$

The strength of the neighbor arm preferences are updated according to the rule:

$$b^{n_i} \Leftarrow b^{n_i}\,(1-\omega) + \alpha\delta_t \times \begin{cases} 1 - p(a = \{a_t + i, a_t - i\}; s), a = a_t \pm i \\ -p(a = \{a_t + i, a_t - i\}; s), a \neq a_t \pm i \end{cases} \tag{5}$$

where i is either 1 or 2 depending on whether the energy being calculated is the immediate neighbor preference or the 2 arm away preference. And, finally, the state-value approximation is updated according to the rule:

$$V(s) \Leftarrow V(s)(1-\omega) + \alpha\delta_t \times \begin{cases} 1, s = s_t \\ 0, s \neq s_t \end{cases} \tag{6}$$

The learning, α, and forgetting, ω, rates were the same for all of the updating rules. This does not need to be the case, but since we found that a single learning and forgetting rate fit the data well, we did not feel there was a need to increase the complexity of the models by increasing the number of parameters.

Model fitting: The model was implemented in C++ and run and fit within Igor Pro (Wavemetrics). There were 7 arms at which the agent could be located, 6 track arms and 1 rest box "arm;" whereas, there we only 6 arms to which the agent could transition. That means that the model implemented the transition from the rest box to the track but did not model the return to the rest box from the track, this was done so that all track arm visits would be included in the analyses. For the working memory version of the model, there were, therefore, 43 states in which the agent could find itself. 36 states (6^2) for all combinations for both the previous and current arm being one of the 6 track arms (6 of these states could never be visited since a return to the same arm is not allowed), an additional 6 states for the current arm being one of the 6 track arms and the previous "arm" being the rest box, and a final 1 state for the agent starting from the rest box.

We fit the various agents to individual animals in two ways. The first was using maximum likelihood; whereby we optimized the parameters of the agent to maximize the likelihood that the agent would exactly reproduce the arm visits of the animal during the second and third spatial alternation contingencies. Given that this method did not find parameters that would generate the behavior with similar statistics to the animals (Fig 2), we use the alternative Approximate Bayesian Computation method. For that fitting we found the parameters that minimized the average root mean square difference between the inbound and outbound errors of the individual animal and of the average of 200 different repeats of the model. The inbound and outbound fitting errors were summed with equal weighting to create the final fitting error. For both fitting categories we used simulated annealing and ran the optimization at least 4 different times from different initial conditions. We chose the parameters with the maximum likelihood or minimal error. For each run of the model we used the same random number generating seed to minimize the random fluctuations between parameter sets (Daw, 2011).

Quantification of differences in statistics between high and low performing rats: We quantified the differences between the high and low performing rats in the neighbor frequency and directional inertia metrics by calculating the average session whereby these metrics dropped to half their range during the first contingency. For the neighbor frequency we defined the range as the difference between the value of the average across all animals at the first and last session of the contingency. For the direction inertia, since it shows more biphasic dynamics, we defined the range as the difference between the value of the average across all animals during the first session of the contingency and the minimal average value of the directional inertia. We permuted the identity of the rats and recalculated the metric 10,000 times to compare to the differences between the high and low performing rats.

Acknowledgements: We thank E. Vértes for helpful discussion and A.K. Gillespie for technical assistance. This work was supported by grants from the Jane Coffin Childs Memorial Fund for Medical Research (D.B.K.), the UCSF Physician Scientist Scholars Program (D.B.K.), an NIH R25 (R25MH060482) (D.B.K), the Simons Foundation for Autism Research grant (291584) (L.M.F.), the Howard Hughes Medical Institute (L.M.F.) and the Max Planck Society (P.D.).

Author Contributions. D.B.K., L.M.F. and P.D. designed the study, D.B.K. and E.A.M. collected the data, D.B.K. and P.D. developed the models, D.B.K. and Z.Y. analyzed the data, D.B.K. designed the automated behavior system, D.K.M. and D.F.L. developed the data acquisition system, and D.B.K., L.M.F. and P.D. wrote the manuscript.

Declaration of interests. The authors declare no competing interests.

Figure Legends

Fig 1. Automated behavior system for analysis of continuous spatial alternation behavior. (A) Layout of automated behavior system. (B) Arm preferences of all rats (n = 24) during the exploratory period of the behavior, where a rat can get rewarded at any arm of the track. Rats ordered by arm preference. (C) Example transition matrix during the exploratory period of the behavior a single rat (animal #8 from A) showing the probability of going to any of the six arms when starting from each of the six arms. (D) Example arm choices (arrowheads) of a single rat (animal #1 from A) during a session of the exploratory behavior. (E) Probability of seeing the maximal arm preference (left) neighbor visit frequency (middle) or directional inertia (right) given random choices between the six arms. Horizontal line shows a probability of 0.05. For the neighbor transition frequency, the random distribution was defined by the arm visit probability of the animal, and for the directional inertia the random distribution was defined by the transition matrix of the animal. As the p-value was determined using 10,000 draws from distributions, the minimal value is 10^{-4}. 14/24 rats were at that minimal value for the max arm probability, 24/24 for the neighbor visit frequency, and 19/24 for the directional inertia. (F) Probability of getting a reward for all 24 rats. Within each contingency, curves smoothed with a Gaussian filter with a standard deviation of 10 arm visits. Two different rats shown in colors (yellow and teal) to indicate consistency of performance in those rats across the different contingencies. Contingencies demarcated by vertical lines above the plot. Contingencies labeled by six boxes, with the open boxes indicating which of the six rewards well contain the potential for reward. (G) Error likelihoods for inbound and outbound trials for all 24 animals. Values smoothed with a Gaussian filter with a standard deviation of 10 inbound or outbound trials and then interpolated to reflect total arm visits. Colors indicate the same rats as in F. Contingencies indicated as in F.

Fig 2. Simple RL agent does not learn alternation task. (A) Graphic of RL agent M1. Colored symbols, $b(a, s_t)$ and V, indicate the entities that change as the agent goes to arms, a, and does or does not get reward, r. The state of this agent, and therefore the probability of transitioning to each of the arms, $p(a; s)$, is defined by the current arm location, a_t, of the agent. (B) Average reward probability of all animals (n = 24) across all contingencies (grey), and average behavior of 200 repeats of the model, whose parameters were optimized to maximize the rewards gotten across all contingencies (purple). (C) Inbound and outbound error likelihood for an individual animal across all contingencies. Grey color indicates animal behavior to which the model is not fit, while black color indicates animal behavior to which the model is fit. Values smoothed with a Gaussian filter with a standard deviation of 2.25 errors and then interpolated to reflect arm visits. In light purple is the average behavior of 200 repeats of the model using the parameters that maximize the likelihood of exactly reproducing the choices of a single run of the model using the parameters that maximize the amount of reward. In yellow is the average behavior of 200 repeats of the model using the parameters that maximize the likelihood of exactly reproducing the rat's arm choices during the second and third alternation contingencies (black). In dark purple is the average behavior of 200 repeats of the model using the parameters that minimize the rms difference between the model and the animal during the second and third alternation contingencies (black). (D) Comparison of parameters under

different fitting conditions shown in C. The leftmost point shows the parameters for the model that maximizes the amount of reward the model can get over all the contingencies, as shown in B.

Fig 3. Working memory RL agent learns alternation task far slower than rats. (A) Graphic of RL agent M2. Colored symbols, $b(a, s_t)$ and V, indicate the entities that change as the agent goes to arms, a, and does or does not get reward, r. The state of this agent, and therefore the probability of transitioning to each of the arms, $p(a; s)$, is defined by the current arm location, a_t, and the previous arm location, a_{t-1}, of the agent. (B) Average reward probability of all animals (n = 24) across all contingencies (grey), and average behavior of 200 repeats of the model with parameters chosen to maximize the rewards received across all contingencies (orange). The model was given extra arm visits to reach asymptotic behavior (after the endpoints of the grey curves for each contingency) to show more clearly the model's ability to learn the task. (C) Inbound and outbound error likelihood for an individual animal across all contingencies (grey/black). Values smoothed with a Gaussian filter with a standard deviation of 2.25 errors and then interpolated to reflect arm visits. In orange is the average behavior of 200 repeats of the model using the parameters that minimize the rms difference between the model and the animal during the second and third alternation contingencies (black). (D) Comparison of the parameters for the fits of individual animals (open circles) to the parameters that maximize rewards (closed circles).

Fig 4. Giving RL agent dynamic arm bias allows for far more rapid learning of alternation task. (A) Graphic of RL agent M3. Colored symbols, $b(a, s_t)$, $b^i(a)$ and V, indicate the entities that change as the agent goes to arms, a, and does or does not get reward, r. The probability of transitioning to each of the arms, $p(a; s)$, is determined by the state of the agent and an independent dynamic arm preference. (B) Average reward probability of all animals (n = 24) across all contingencies (grey), and average behavior of 200 repeats of the model, whose parameters were optimized to maximize the rewards gotten across all contingencies (blue). (C) Inbound and outbound error likelihood for an individual animal across all contingencies (grey/black). Values smoothed with a Gaussian filter with a standard deviation of 2.25 errors and then interpolated to reflect arm visits. In blue is the average behavior of 200 repeats of the model using the parameters that minimize the rms difference between the model and the animal during the second and third alternation contingencies (black). (D) Comparison of the parameters for the fits of individual animals (open circles) to the parameters that maximize rewards (closed circles).

Fig 5. Adding dynamic neighbor bias enables further speeding up of RL agent learning of alternation task. (A) Graphic of RL agent M4. Colored symbols, $b(a, s_t)$, $b^i(a)$, $b^{n_1}\chi_1(a = a_t \pm 1)$, $b^{n_2}\chi_2(a = a_t \pm 2)$and V, indicate the entities that change as the agent goes to arms, a, and does or does not get reward, r. The probability of transitioning to each of the arms, $p(a|s)$, is determined by the state of the agent, an independent dynamic arm preference, and dynamic

neighbor preferences. (B) Average reward probability of all animals (n = 24) across all contingencies (grey), and average behavior of 200 repeats of the model, whose parameters were optimized to maximize the rewards gotten across all contingencies (green). (C) Inbound and outbound error likelihood for an individual animal across all contingencies (grey/black). Values smoothed with a Gaussian filter with a standard deviation of 2.25 errors and then interpolated to reflect arm visits. In green is the average behavior of 200 repeats of the model using the parameters that minimize the rms difference between the model and the animal during the second and third alternation contingencies (black). (D) Comparison of the parameters for the fits of individual animals (open circles) to the parameters that maximize rewards (closed circles). (E) RMS difference between the model and the data for all animals (n = 24) for the inbound and outbound errors for each contingency for the different models from Fig 3 (orange), 4 (blue) and 5 (green).

Fig 6. Successes and failures of the model. (A) Inbound and outbound error likelihood for an individual animal across all contingencies (grey/black). Values smoothed with a Gaussian filter with a standard deviation of 2.25 errors and then interpolated to reflect arm visits. In dark green is the average behavior of 200 repeats of the model using the parameters that minimize the rms difference between the model and the animal during the second and third alternation contingencies (black). In light green is the inbound and outbound error likelihood for a single repeat of the model using the parameters from the fit to the individual animal. (B) Cumulative distribution for all of the RMS difference errors between individual simulations of the model and the 200 repeats of the model for the outbound errors of the fourth contingency. The RMS difference error for the data is shown in the black circle. Fit accuracy in the subsequent panel is the fraction of the cumulative distribution that falls to the right of the data. (C) Histograms across animals of the log of the fit accuracy of the model from Fig 5. Above each pair of histograms are boxes indicating the contingency. Open boxes indicate wells with the potential to provide reward. Fit accuracy is the probability of individual runs of the model providing an error greater than the error of the data. Parameters optimized to minimize the error between 200 repeats of the model and the data during the second and third alternation contingency (black). Vertical dotted line show $p = 0.05$ for comparison. Arrowheads point to value of the median for the outbound fit accuracies. (D) Difference between the error likelihood for the rats and the model fit to the individual rats, averaged across all rats (± standard deviation). Positive residual values indicate that the model had higher error likelihoods and negative residual values indicate that the model had lower error likelihoods.

Fig 7. M4 captures differences in animal performance. (A) Three-dimensional space of parameters projected down on all pairs of parameters. The median and interquartile range of 24 fits for each animal is plotted as the red or pink dot with errors bars in both axes. Red and pink colors indicate split of animals as dictated by panel B. Color scale in background is the maximal reward rate during the second and third contingency for the pair of parameters. For instance, for the α/γ plot, each color indicates the maximal reward value that can be obtained for the pairing of α and γ, which is found by scanning through all values of ω. (B) The reward

obtained by each animal (diamonds) and by the model fit to each animal (circle). Animals ordered by the model reward. Animals colored by the median split for the actual reward rate (blue/black) and model reward rate (pink/red). Dotted horizontal line indicates maximal reward ratefor the model. (C) Average (± sem) probability correct across all contingencies for the grouping by the median split of the actual reward rate. (D) Average (± sem) probability correct across all contingencies for the grouping by the median split of the model reward rate. (E) Difference between the average reward rates (± sem) when the animals are split based upon their actual reward rates (black) or based upon the model reward rates (red).

Fig 8. Comparison of behavioral statistics to understand differences in the first contingency. Neighbor arm visit frequency (A), and directional inertia (B) for each session during the exploratory period and first spatial alternation contingency. In grey is the average (± sem) behavior of all 24 rats. In orange (green) is the average (± sem) behavior of the 24 fits of model M2 (M4). The average (± sem) neighbor arm visit frequency (C), and directional inertia (D) for the higher (red) and lower performing (pink) rats as defined by the median split of the model reward rate in Fig 7B. The horizontal solid lines in C & D indicates the values at which the decay (C) or rise (D) are calculated to measure differences between the groups. For A – D, the horizontal dotted lines indicate the value for perfect performance of the first alternation contingency. (E) Correlation between the average neighbor frequency during exploration and the average reward rate during the first contingency across all animals. Dotted line shows the linear fit to the data.

Supplementary Figure 1. Reward well. (A) Picture of a reward well showing the spigot through which milk is delivered, flanked by an IR LED and phototransister, encased in metal elbows, to detect the position of the animal. Any unconsumed milk exited the track through the drain below the spigot. A light is illuminated directly behind the reward well when there is potential for reward delivery (see methods). Reward wells were made entirely out of metal.

Supplementary Figure 2. Males and females perform the behavior comparably across the two cohorts. (A) Average probability of getting a reward for the male (dotted line) and female (solid line) rats in the first (top) and second (bottom) cohort. Within each contingency, curves smoothed with a Gaussian filter with a standard deviation of 10 arm visits and then averaged across the different animals. Thickness of the line indicates the sem. Contingencies demarcated by vertical lines above the plot. Contingencies indicated by filled in boxes indicating the potential for reward at each of the six reward wells. (B) Average missed poke likelihood across all contingencies. Averaged across all rats. Thickness of line indicates sem. (C) Maximal reward rate in a session for each contingency and for all animals. Horizontal dotted line demarcates 80% correct.

Supplementary Figure 3. Dynamics of individual arm and neighbor preferences for the model. The average individual arm preferences (β^{a}) (A) and neighbor arm preferences (β^{n_1} and β^{n_2}) across all contingencies and repeats of the model for the fit to the animal shown in Fig 5C. The values shown are those prior to passing through the exponential for the Softmax.

Supplementary Figure 4. Error landscape for M4. (A) Three-dimensional space of parameters projected onto the plane for the parameters from the fit. For instance, for the α/γ plot, the plane for the fit value of ω is chosen. The median and interquartile range for the parameters for the same rat from Figs 2- 6 for 24 fits are plotted as the red dot with errors bars in both axes (obscured by the dot). The color scale in the background shows the error between the model and the data.

Brunton, B.W., Botvinick, M.M., Brody, C.D., 2013. Rats and humans can optimally accumulate evidence for decision-making. Science 340, 95–98. doi:10.1126/science.1233912

Daw, N.D., 2011. Trial-by-trial data analysis using computational models, in: Decision Making, Affect, and Learning: Attention and Performance XXIII. Oxford University Press, pp. 3–38. doi:10.1093/acprof:oso/9780199600434.003.0001

Dudchenko, P.A., 2004. An overview of the tasks used to test working memory in rodents. Neurosci Biobehav Rev 28, 699–709. doi:10.1016/j.neubiorev.2004.09.002

Fernandez-Ruiz, A., Oliva, A., de Oliveira, E.F., Rocha-Almeida, F., Tingley, D., Buzsáki, G., 2019. Long-duration hippocampal sharp wave ripples improve memory. Science 364, 1082–. doi:10.1126/science.aax0758

Flagel, S.B., Clark, J.J., Robinson, T.E., Mayo, L., Czuj, A., Willuhn, I., Akers, C.A., Clinton, S.M., Phillips, P.E.M., Akil, H., 2011. A selective role for dopamine in stimulus-reward learning. Nature 469, 53–57. doi:10.1038/nature09588

Flagel, S.B., Robinson, T.E., Clark, J.J., Clinton, S.M., Watson, S.J., Seeman, P., Phillips, P.E.M., Akil, H., 2010. An animal model of genetic vulnerability to behavioral disinhibition and responsiveness to reward-related cues: implications for addiction. Neuropsychopharmacology 35, 388–400. doi:10.1038/npp.2009.142

Frank, L.M., Stanley, G.B., Brown, E.N., 2004. Hippocampal plasticity across multiple days of exposure to novel environments. Journal of Neuroscience 24, 7681–7689. doi:10.1523/JNEUROSCI.1958-04.2004

Galsworthy, M.J., Paya-Cano, J.L., Monleon, S., Plomin, R., 2002. Evidence for general cognitive ability (g) in heterogeneous stock mice and an analysis of potential confounds. Genes Brain Behav. 1, 88–95.

Gershman, S.J., Daw, N.D., 2017. Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. Annual review of psychology 68, 101–128. doi:10.1146/annurev-psych-122414-033625

Gershman, S.J., Niv, Y., 2010. Learning latent structure: carving nature at its joints. Curr Opin Neurobiol 20, 251–256. doi:10.1016/j.conb.2010.02.008

International Brain Laboratory, 2017. An International Laboratory for Systems and Computational Neuroscience. Neuron 96, 1213–1218. doi:10.1016/j.neuron.2017.12.013

Jadhav, S.P., Kemere, C., German, P.W., Frank, L.M., 2012. Awake hippocampal sharp-wave ripples support spatial memory. Science 336, 1454–1458. doi:10.1126/science.1217230

Kaelbling, L.P., Littman, M.L., Cassandra, A.R., 1998. Planning and acting in partially observable stochastic domains. Artificial Intelligence 101, 99–134. doi:10.1016/S0004-3702(98)00023-X

Karlsson, M.P., Frank, L.M., 2008. Network dynamics underlying the formation of sparse, informative representations in the hippocampus. Journal of Neuroscience 28, 14271–14281. doi:10.1523/JNEUROSCI.4261-08.2008

Kim, S.M., Frank, L.M., 2009. Hippocampal lesions impair rapid learning of a continuous spatial alternation task. PLoS One 4, e5494. doi:10.1371/journal.pone.0005494.t001

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R., 2017. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci USA 114, 3521–3526. doi:10.1073/pnas.1611835114

Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J., 2017. Building machines that learn and think like people. Behav Brain Sci 40, e253. doi:10.1017/S0140525X16001837

Li, J., Daw, N.D., 2011. Signals in human striatum are appropriate for policy update rather than value prediction. Journal of Neuroscience 31, 5504–5511. doi:10.1523/JNEUROSCI.6316-10.2011

Lintusaari, J., Gutmann, M.U., Dutta, R., Kaski, S., Corander, J., 2017. Fundamentals and Recent Developments in Approximate Bayesian Computation. Syst. Biol. 66, e66–e82. doi:10.1093/sysbio/syw077

Lloyd, K., Becker, N., Jones, M.W., Bogacz, R., 2012. Learning to use working memory: a reinforcement learning gating model of rule acquisition in rats. Front Comput Neurosc 6, 87. doi:10.3389/fncom.2012.00087

Luksys, G., Gerstner, W., Sandi, C., 2009. Stress, genotype and norepinephrine in the prediction of mouse behavior using reinforcement learning. Nat Neurosci 12, 1180–1186. doi:10.1038/nn.2374

Matzel, L.D., Han, Y.R., Grossman, H., Karnik, M.S., Patel, D., Scott, N., Specht, S.M., Gandhi, C.C., 2003. Individual differences in the expression of a “general” learning ability in mice. J Neurosci 23, 6423–6433.

Matzel, L.D., Townsend, D.A., Grossman, H., Han, Y.R., Hale, G., Zappulla, M., Light, K., Kolata, S., 2006. Exploration in outbred mice covaries with general learning abilities irrespective of stress reactivity, emotionality, and physical attributes. Neurobiol Learn Mem 86, 228–240. doi:10.1016/j.nlm.2006.03.004

Packard, M.G., Goodman, J., 2013. Factors that influence the relative use of multiple memory systems. Hippocampus 23, 1044–1052. doi:10.1002/hipo.22178

Packard, M.G., McGaugh, J.L., 1996. Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. Neurobiol Learn Mem 65, 65–72. doi:10.1006/nlme.1996.0007

Poddar, R., Kawai, R., Olveczky, B.P., 2013. A fully automated high-throughput training system for rodents. PLoS One 8, e83171. doi:10.1371/journal.pone.0083171

Rivalan, M., Munawar, H., Fuchs, A., Winter, Y., 2017. An Automated, Experimenter-Free Method for the Standardised, Operant Cognitive Testing of Rats. PLoS One 12, e0169476. doi:10.1371/journal.pone.0169476

Sigurdsson, T., Stark, K.L., Karayiorgou, M., Gogos, J.A., Gordon, J.A., 2010. Impaired hippocampal-prefrontal synchrony in a genetic mouse model of schizophrenia. Nature 464, 763–767. doi:10.1038/nature08855

Singer, A.C., Frank, L.M., 2009. Rewarded outcomes enhance reactivation of experience in the hippocampus. Neuron 64, 910–921. doi:10.1016/j.neuron.2009.11.016

Singer, A.C., Karlsson, M.P., Nathe, A.R., Carr, M.F., Frank, L.M., 2010. Experience-dependent development of coordinated hippocampal spatial activity representing the similarity of related locations. Journal of Neuroscience 30, 11586–11604. doi:10.1523/JNEUROSCI.0926-10.2010

Sorge, R.E., Martin, L.J., Isbester, K.A., Sotocinal, S.G., Rosen, S., Tuttle, A.H., Wieskopf, J.S., Acland, E.L., Dokova, A., Kadoura, B., Leger, P., Mapplebeck, J.C.S., McPhail, M., Delaney, A., Wigerblad, G., Schumann, A.P., Quinn, T., Frasnelli, J., Svensson, C.I., Sternberg, W.F., Mogil,

J.S., 2014. Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nat Methods 11, 629–632. doi:10.1038/nmeth.2935

Stead, J.D.H., Clinton, S., Neal, C., Schneider, J., Jama, A., Miller, S., Vazquez, D.M., Watson, S.J., Akil, H., 2006. Selective breeding for divergence in novelty-seeking traits: Heritability and enrichment in spontaneous anxiety-related behaviors. Behav. Genet. 36, 697–712. doi:10.1007/s10519-006-9058-7

Suri, R.E., Schultz, W., 1999. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience 91, 871–890. doi:10.1016/s0306-4522(98)00697-6

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning. MIT Press.

Todd, M.T., Niv, Y., Cohen, J.D., 2009. Learning to Use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement, in: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 21. Curran Associates, Inc., pp. 1689–1696.

Williams, R.J., 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning.

Wilson, R., Collins, A., 2019. Ten simple rules for the computational modeling of behavioral data. PsyArXiv.

Zenke, F., Ben Poole, Ganguli, S., 2017. Continual learning through synaptic intelligence 3987–3995.

Zilli, E.A., Hasselmo, M.E., 2008. Modeling the role of working memory and episodic memory in behavioral tasks. Hippocampus 18, 193–209. doi:10.1002/hipo.20382

Figure 1

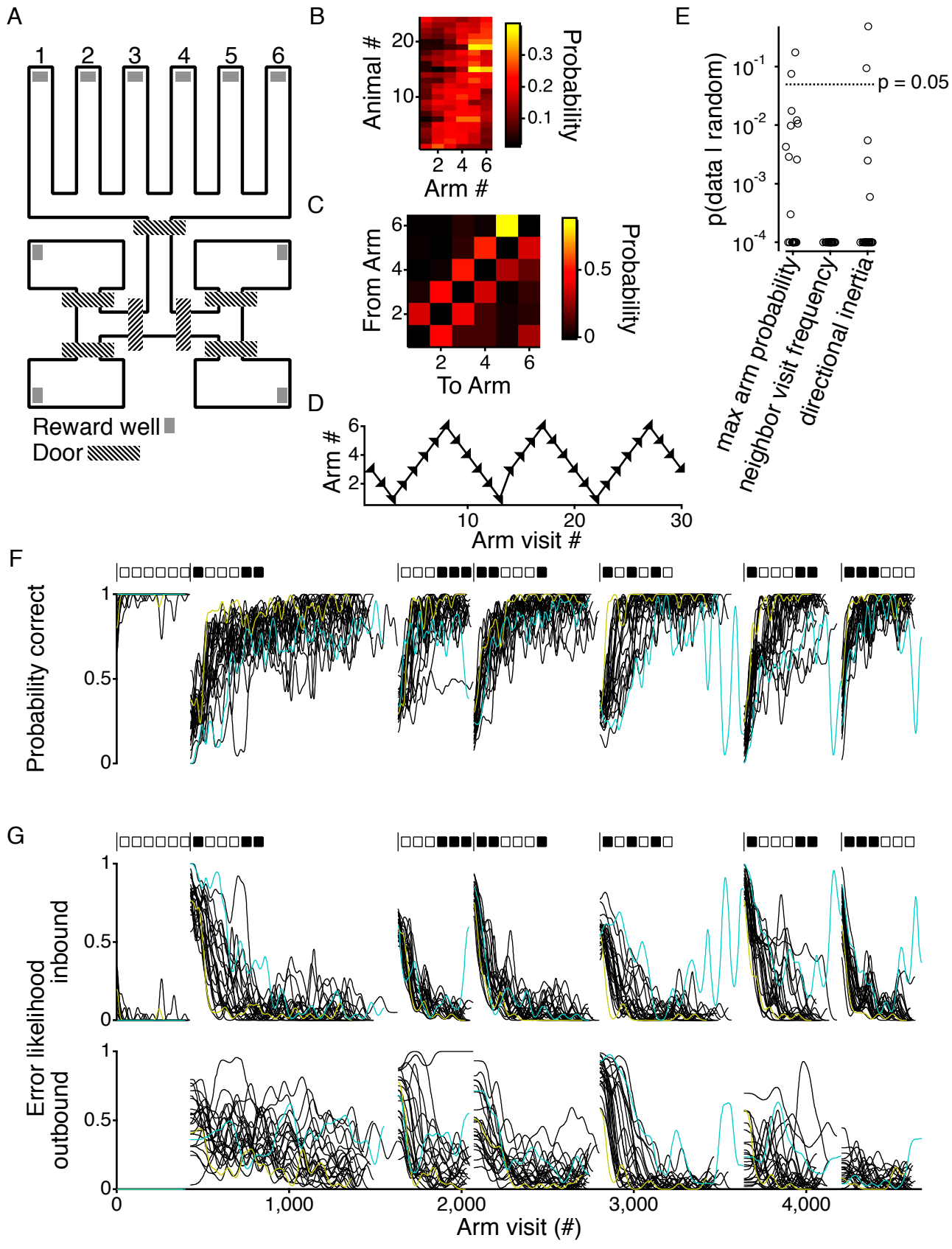

A
1 2 3 4 5 6
Reward well
Door
B
Animal #
Arm #
Probability
C
From Arm
To Arm
Probability
D
Arm #
Arm visit #
E
p(data | random)
p = 0.05
max arm probability
neighbor visit frequency
directional inertia
F
Probability correct
G
Error likelihood
inbound
outbound
Arm visit (#)

Figure 2

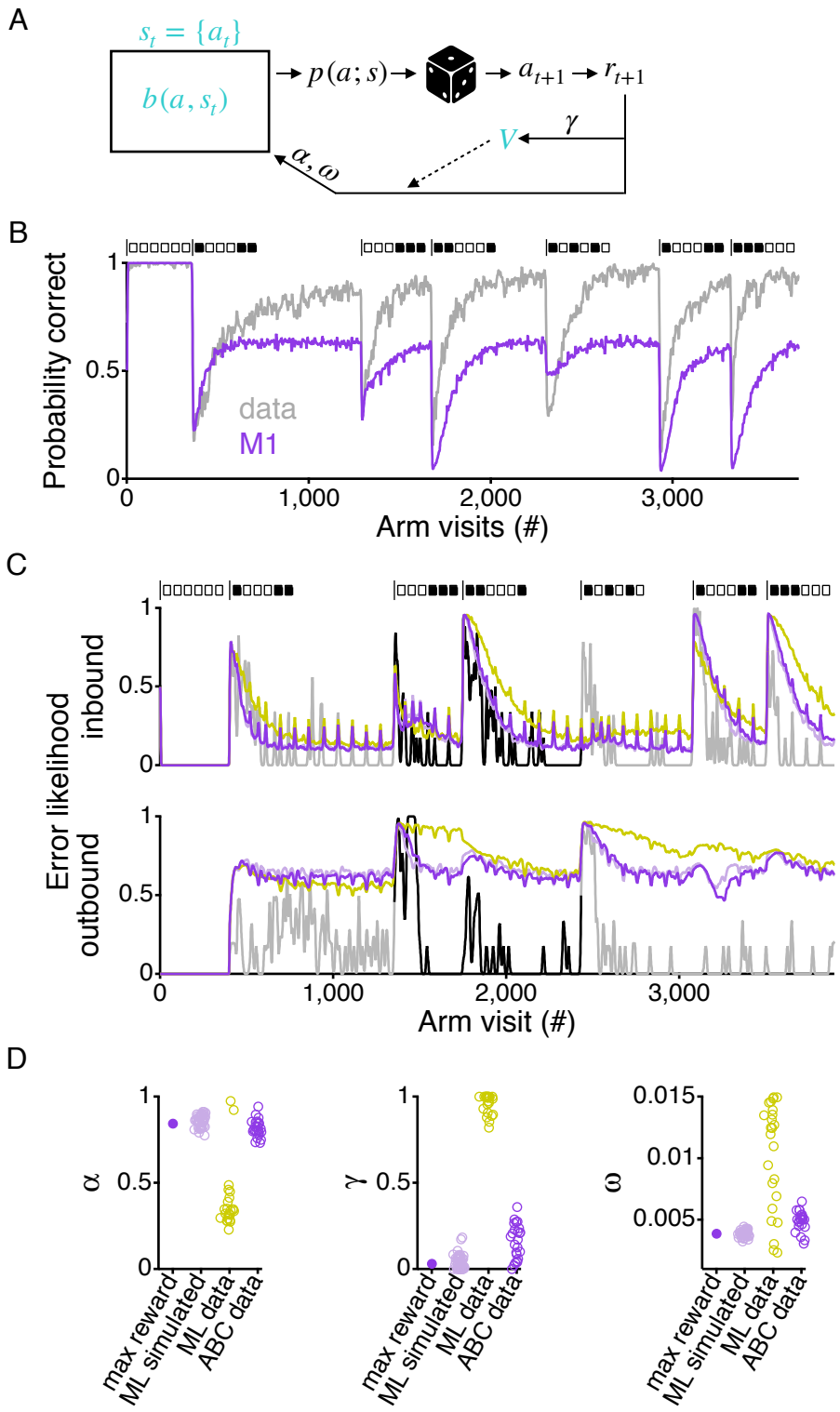

A
$s_t = \{a_t\}$
$b(a, s_t)$
$p(a; s)$
a_{t+1}
r_{t+1}
γ
V
α, ω
B
Probability correct
data
M1
Arm visits (#)
C
Error likelihood
inbound
outbound
Arm visit (#)
D
α
γ
ω
max reward
ML simulated
ML data
ABC data

Figure 3

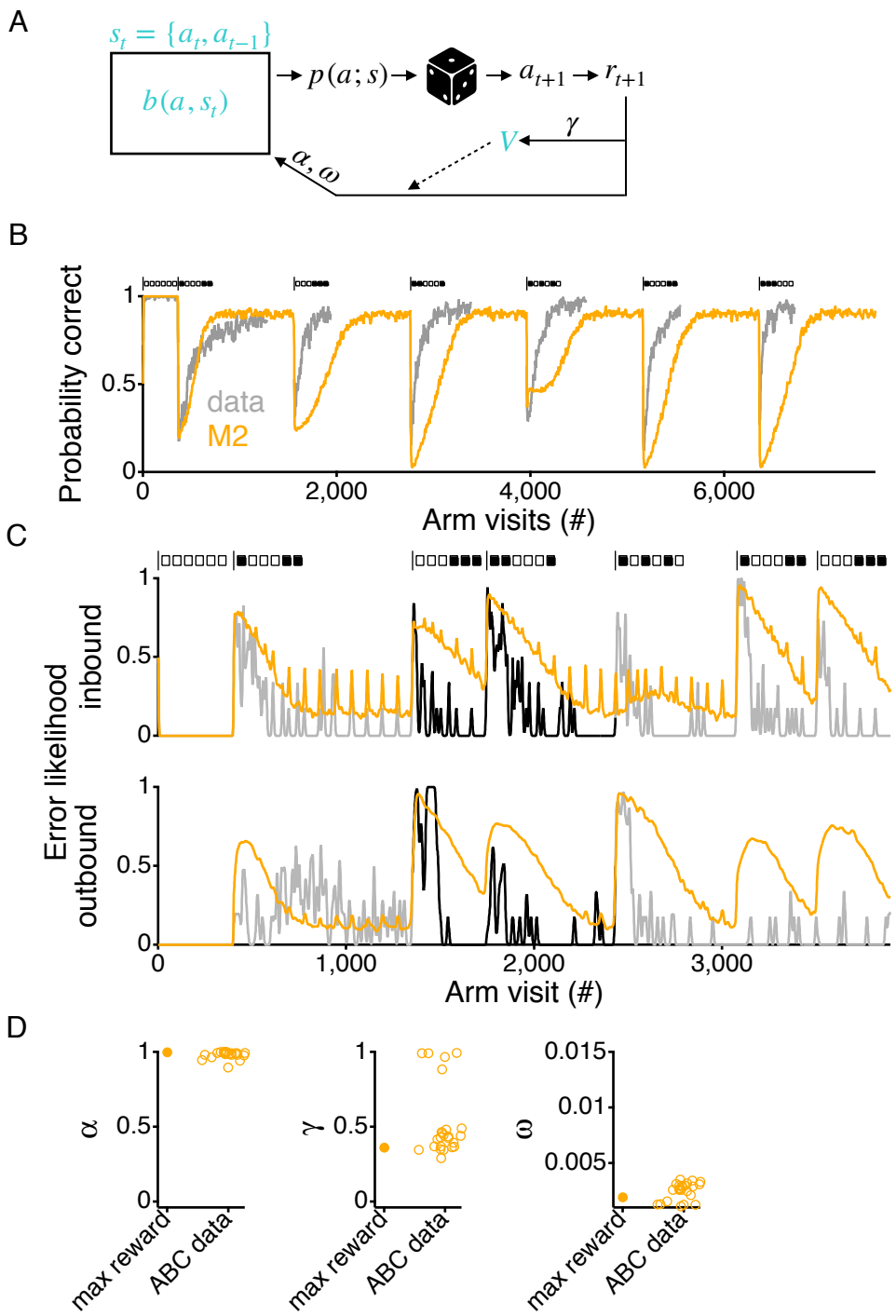

A
$s_t = \{a_t, a_{t-1}\}$
$b(a, s_t)$
$p(a; s)$
a_{t+1}
r_{t+1}
γ
V
α, ω
B
Probability correct
1
0.5
0
data
M2
0
2,000
4,000
6,000
Arm visits (#)
C
Error likelihood
inbound
outbound
1
0.5
0
0
1,000
2,000
3,000
Arm visit (#)
D
α
γ
ω
0.015
0.01
0.005
max reward
ABC data

Figure 4

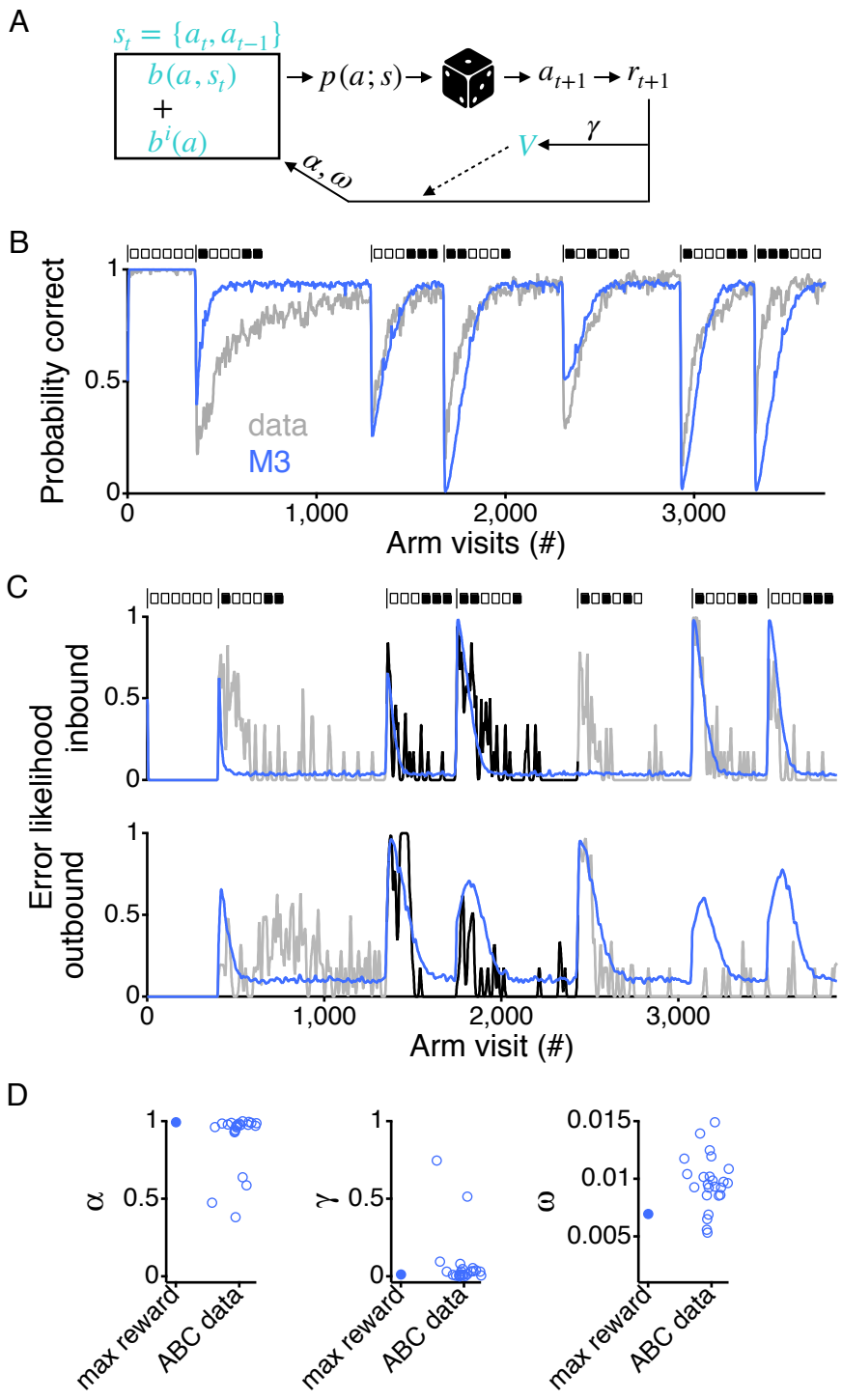

A
$s_t = \{a_t, a_{t-1}\}$
$b(a, s_t)$
+
$b^i(a)$
$p(a; s)$
a_{t+1}
r_{t+1}
γ
V
α, ω
B
Probability correct
data
M3
Arm visits (#)
C
Error likelihood
inbound
outbound
Arm visit (#)
D
α
γ
ω
max reward
ABC data

Figure 5

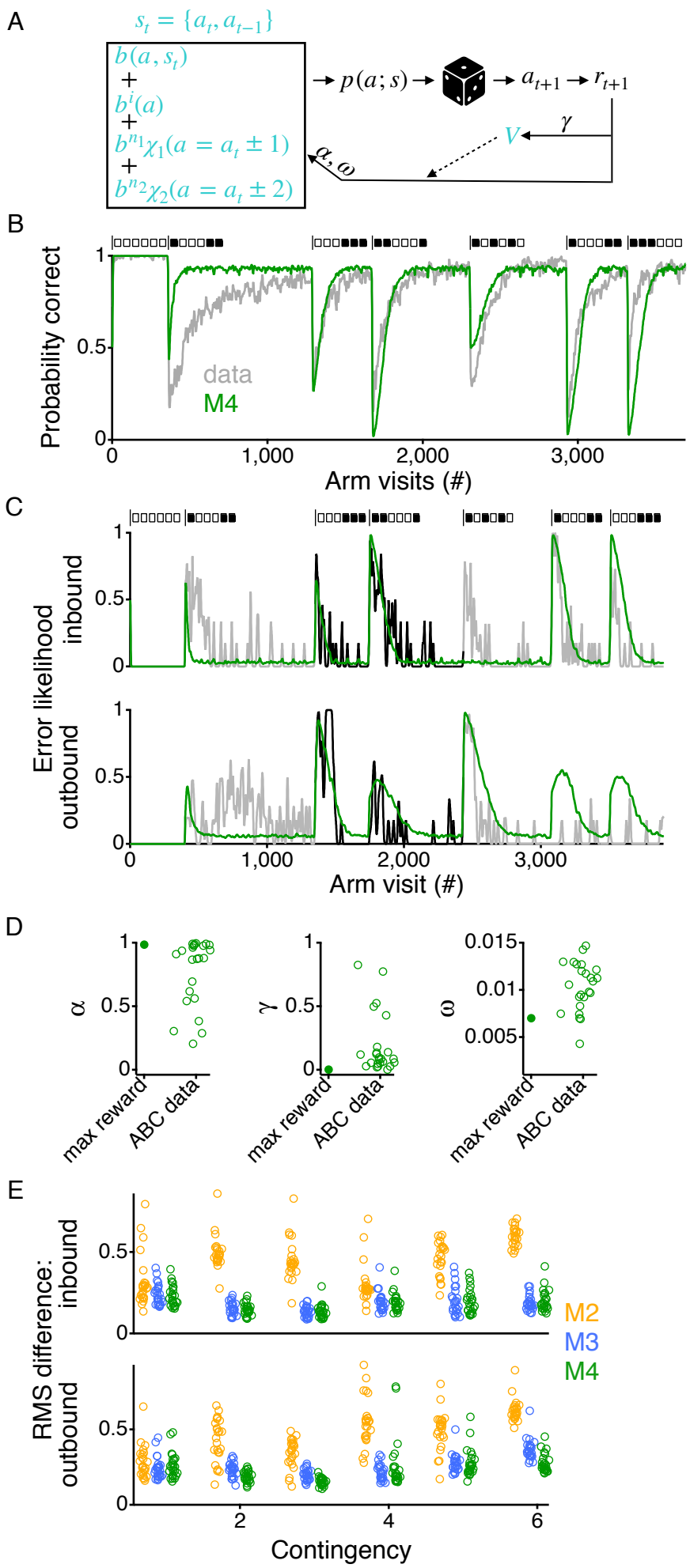

A
$s_t = \{a_t, a_{t-1}\}$
$b(a, s_t)$
+
$b^i(a)$
+
$b^{n_1}\chi_1(a = a_t \pm 1)$
+
$b^{n_2}\chi_2(a = a_t \pm 2)$
$p(a; s)$
a_{t+1}
r_{t+1}
γ
V
α, ω
B
Probability correct
data
M4
Arm visits (#)
C
Error likelihood
inbound
outbound
Arm visit (#)
D
α
γ
ω
max reward
ABC data
E
RMS difference:
inbound
outbound
M2
M3
M4
Contingency

Figure 6

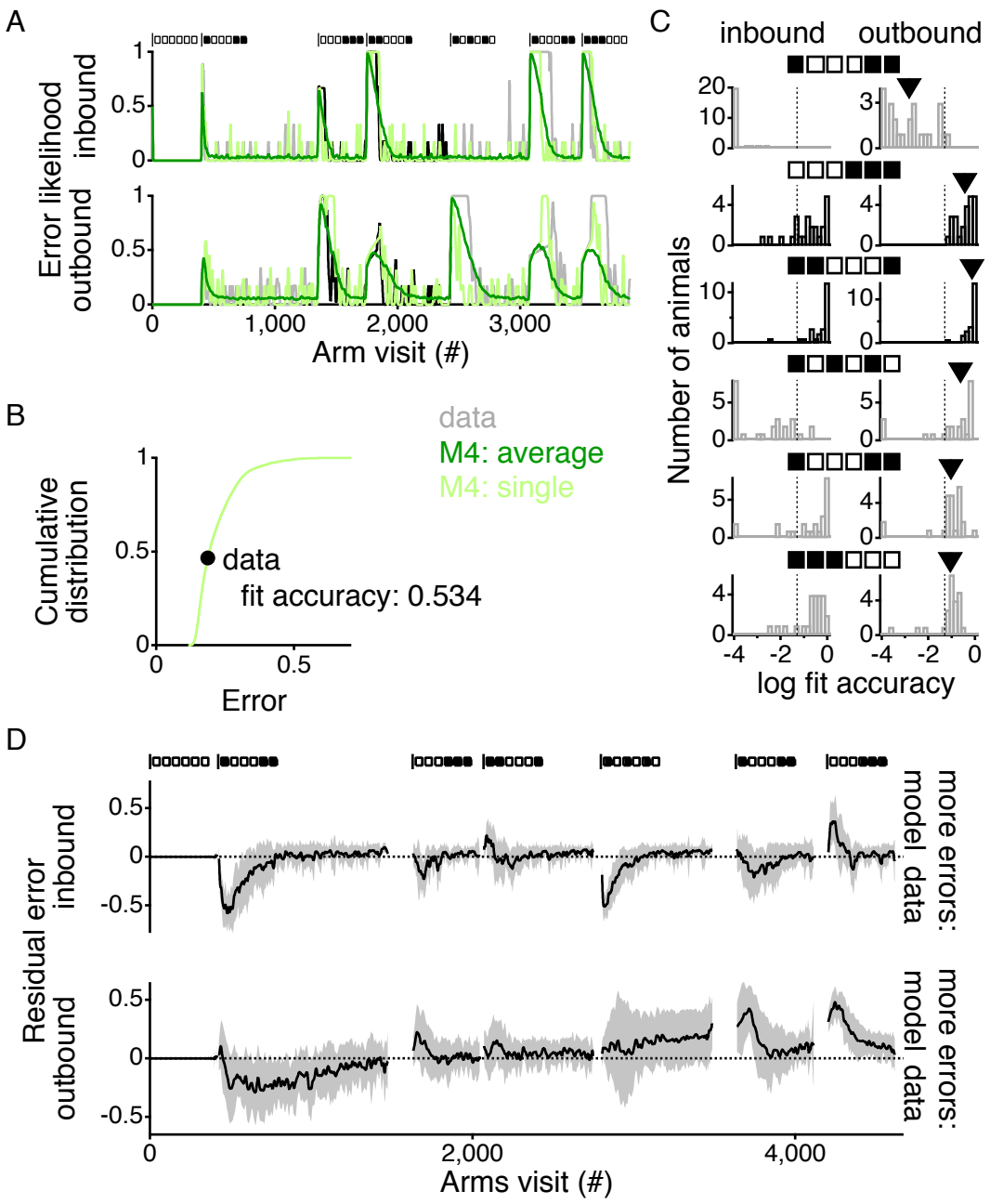

A
Error likelihood
inbound
outbound
Arm visit (#)
B
Cumulative distribution
Error
data
M4: average
M4: single
data
fit accuracy: 0.534
C
inbound
outbound
Number of animals
log fit accuracy
D
Residual error
inbound
outbound
Arms visit (#)
more errors: model data
more errors: model data

Figure 7

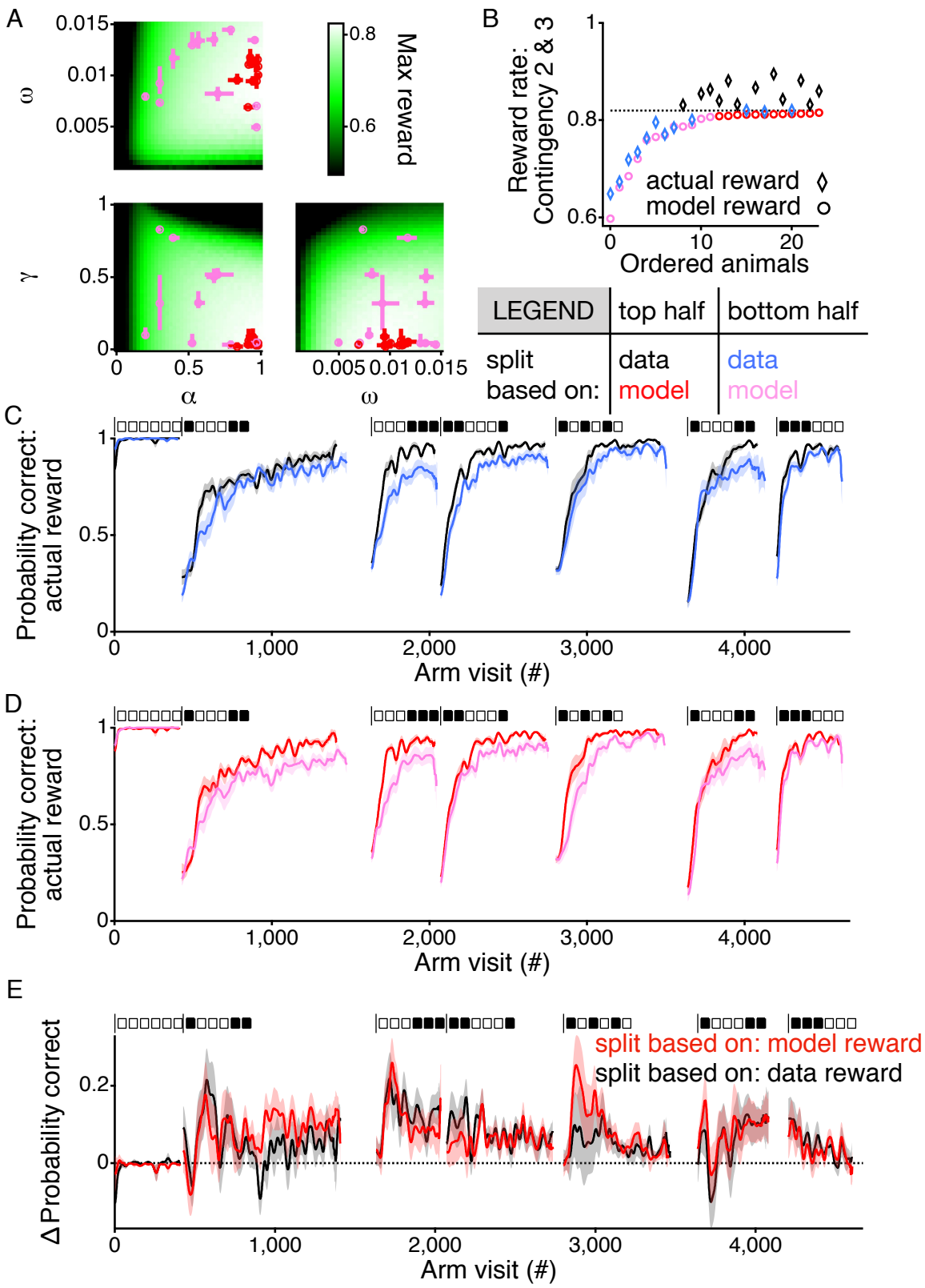
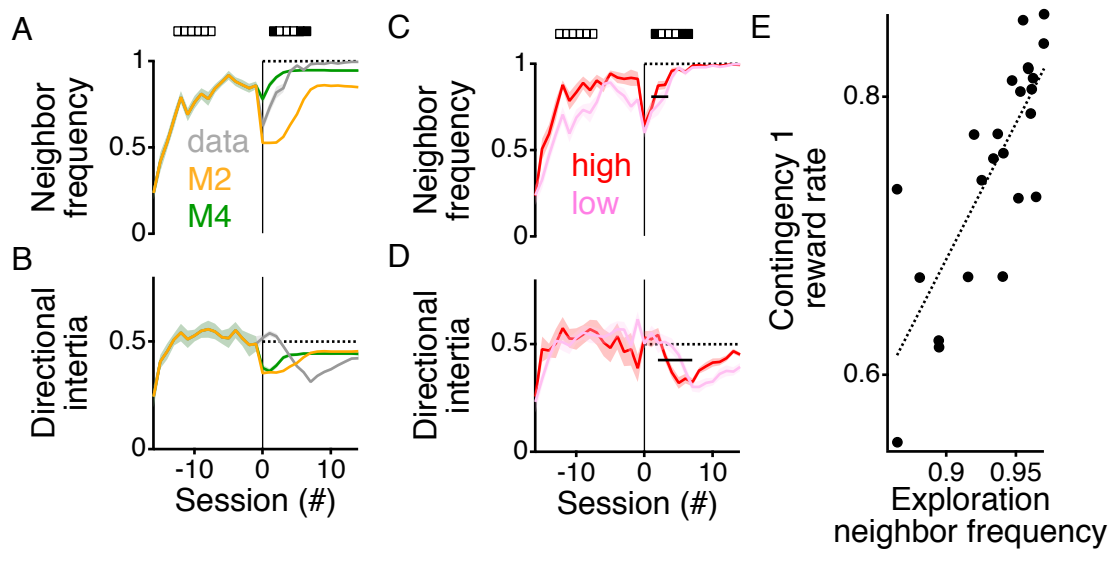
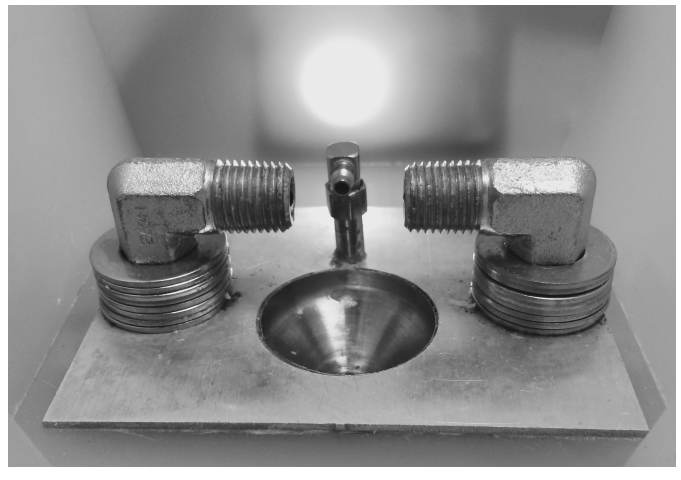

A
ω
γ
α
Max reward
B
Reward rate: Contingency 2 & 3
Ordered animals
actual reward
model reward
LEGEND
top half
bottom half
split based on:
data
model
data
model
C
Probability correct: actual reward
Arm visit (#)
D
Probability correct: actual reward
Arm visit (#)
E
Δ Probability correct
split based on: model reward
split based on: data reward
Arm visit (#)

Figure 8

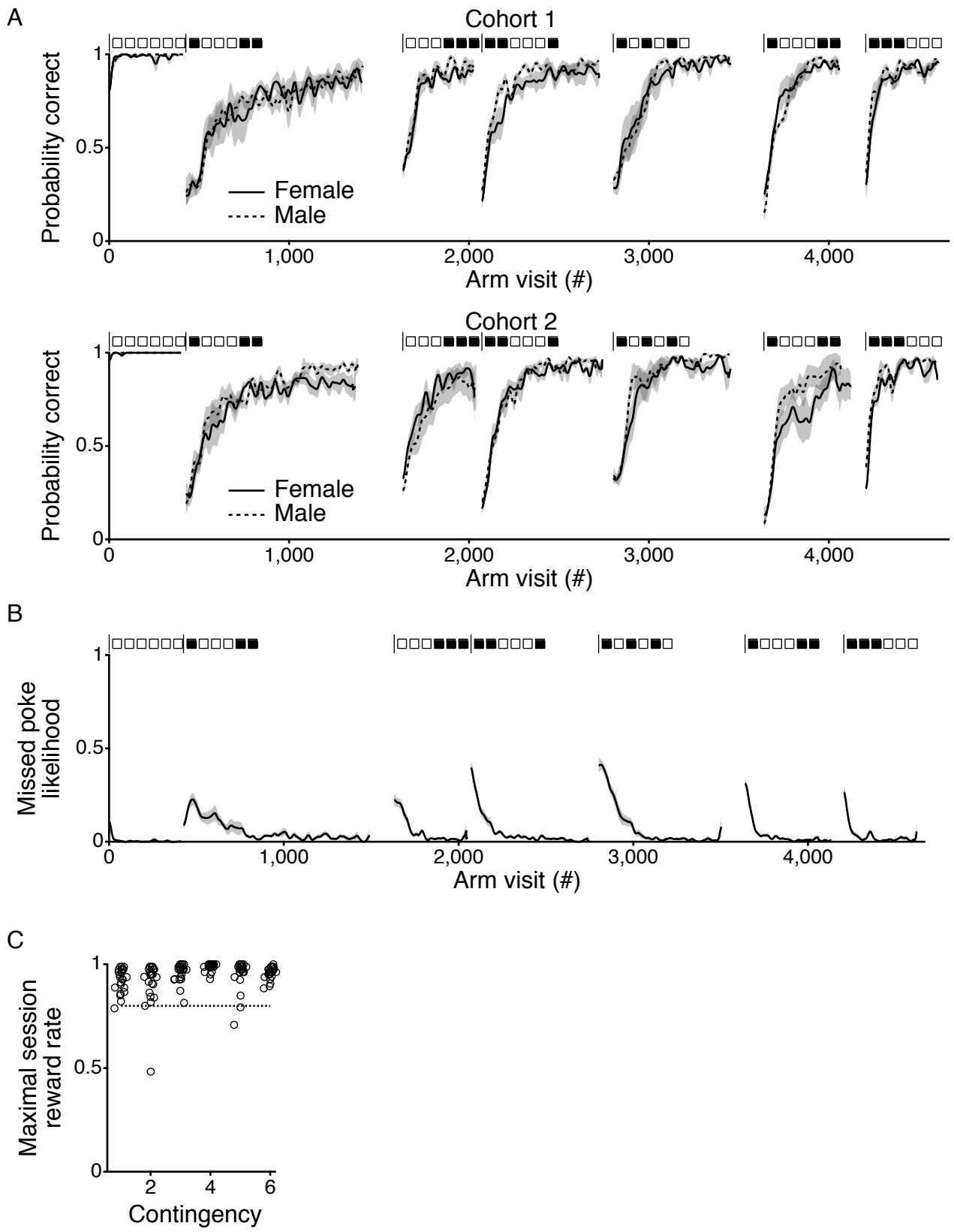

A
B
C
D
E
Neighbor frequency
Directional intertia
data
M2
M4
high
low
Session (#)
Contingency 1 reward rate
Exploration neighbor frequency
1
0.5
0
-10
10
0.8
0.6
0.9
0.95

Supplementary Figure 1

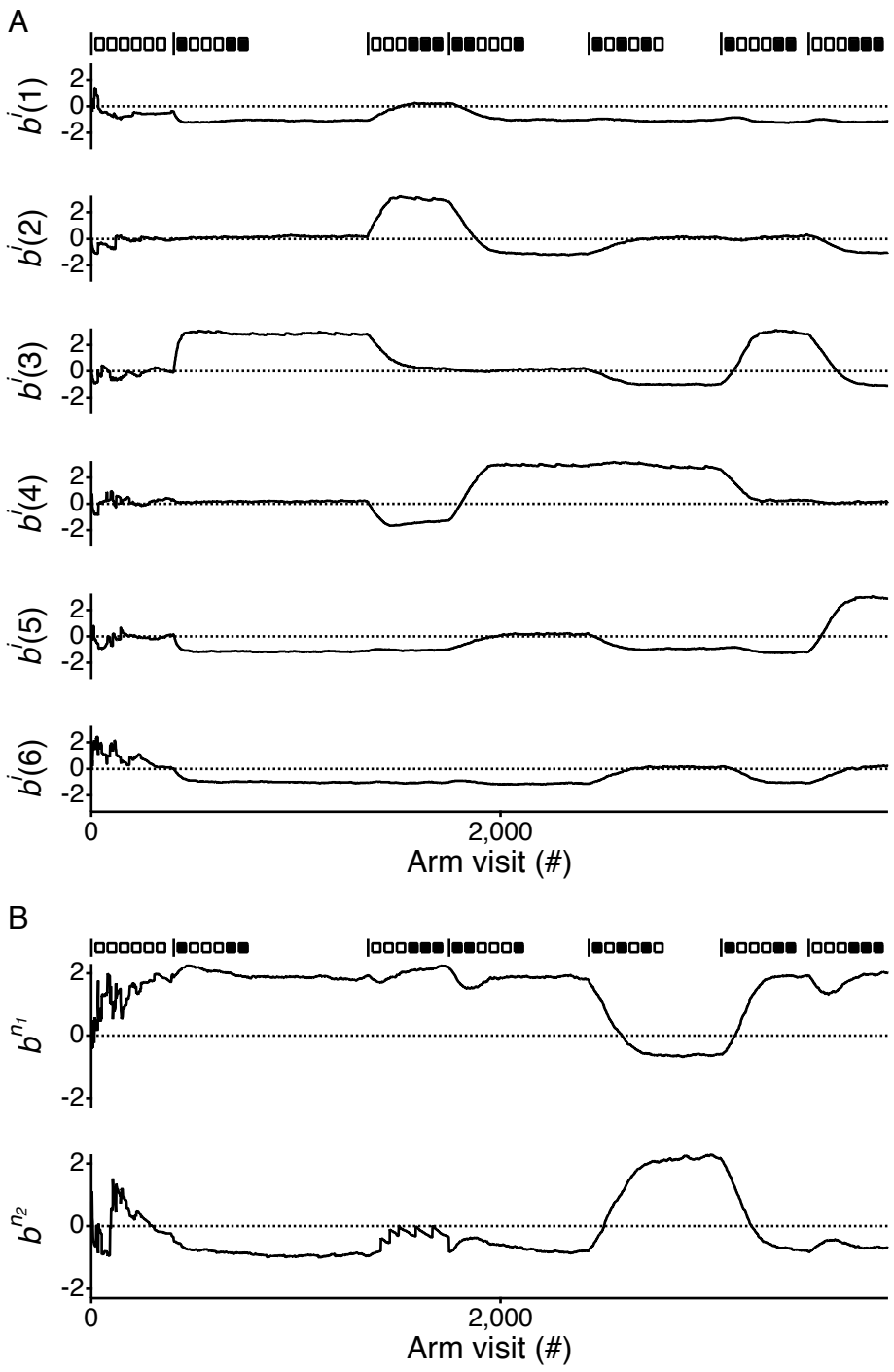
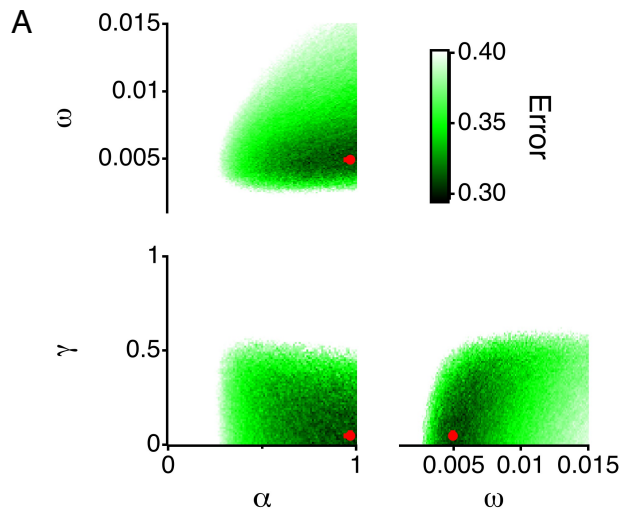
A



Supplementary Figure 2



Supplementary Figure 3



Supplementary Figure 4