# Rapid spatial learning via efficient exploration and inference

Nada Abdelrahman, Wanchen Jiang, Joshua T Dudman*, Ann M Hermundstad*; *correspondence

Animals quickly learn to navigate to rewarding or salient landmarks in their environments. However, existing models often require thousands of trials to learn contingencies that animals learn within tens of trials, and they do so via unstructured sequences of actions that do not mimic real behavior [1]. In this work, we study rapid learning in a hidden-target foraging task for mice [2] in which animals learn to intercept an uncued target location within an open arena. To study the computational underpinnings of this learning, we build an agent that controls its speed and heading over time via a pre-specified set of generative functions; the parameters of these functions can be chosen to smoothly link pairs of spatial locations ("anchor points"). To support learning, we assume that the agent maintains and updates a belief about the target location, which is in turn used to sample anchor points that guide the composition of subsequent trajectories. Three key features enable rapid learning: firstly, learning operates over a low-dimensional set of generative model parameters, rather than a high-dimensional set of discrete location-action pairs; secondly, the agent learns from both rewarded and unrewarded trajectories; lastly, the agent samples anchor points that efficiently narrow down the space of hypothesized target locations by iteratively halving it. As a result, the agent learns within tens of trials to intercept new targets regardless of their spatial separation, matching learning rates observed in mice and significantly outperforming standard reinforcement learning models. In doing so, the agent replicates new features of behavior, such as the progression from more extended to more compact trajectories during learning. Together, this work integrates concepts that have typically been treated separately—such as motor planning, execution, and spatial learning—to understand how animals efficiently explore space and quickly modify their behavior based on experience.

**Methods and Results.** We consider a task in which mice learn to intercept a hidden target in an open spatial arena (Fig. 1A, [2]). Importantly, there are no localizing cues to signal the target location, and animals only receive feedback after returning to a home port to collect reward (Fig. 1B). To study how animals learn to rapidly modify their behavior in this setting, we build on [2] to construct an agent that explores its environment via structured behavioral trajectories, and we formulate learning as the process of finding the parameters of the generative model that enable the agent to intercept the target location. We generate trajectories using parameterized speed and heading functions; given any pair of locations in space ("anchor points"), these functions compute the necessary parameters to generate a trajectory between these locations (Fig. 1C). We assume that the agent knows that there is a single localized target in the arena, and that it maintains a belief about its location. This belief, which we construct in polar coordinates, specifies the probability that the target is at a distance $r^*$ and angle $\theta^*$ from the home port (Fig. 2A). The agent iteratively updates the belief over time based on all past trajectories $\tau_{j\leq i}$ and outcomes $o_{j\leq i}$ (where $i$ indexes the current trial). After running a trajectory $\tau_i$ through a set of locations $\{r, \theta\}_i$ and observing an outcome $o_i$, the agent computes the likelihood of observing the current outcome under different possible target locations, and uses this likelihood to update its belief $P_i(r^*, \theta^*|\tau_{j\leq i}, o_{j\leq i})$ about the target location $(r^*, \theta^*)$. We assume that the agent knows that the target is localized in space, and assumes that the probability of observing a reward falls off with distance from the target location. Thus, if the agent executes a rewarding trajectory, the agent knows that the target is likely to be along the executed path, and unlikely to be far away. We construct this likelihood by convolving a Gaussian kernel with the executed trajectory; after a rewarded trial, this results in localized probability mass around executed anchor points where the agent spent more time (and conversely, after an unrewarded trial, this pushes probability mass away from executed anchor points). The agent uses this belief to sample anchor points for generating new trajectories; the generative model then plans and executes paths between these anchor points (Fig. 2A). We design an efficient sampling process that iteratively narrows down the space of hypothesized target locations by selecting half of the peaks in the posterior belief at which to place anchor points. These anchors guide the generative process in constructing future runs to further localize the target (Fig. 2B). Finally, the posterior enables the agent to evaluate their subjective level of surprise upon receiving outcomes that are inconsistent with their executed trajectory under their current belief; our agent uses this signal to detect putative changes in the environment and reset its prior. Altogether, this enables our agent to find targets within handfuls of trials, irrespective of their location (Fig. 3A,C), mimicking the learning behavior we observe in mice (Fig. 3B,C).

[1] Shamash, Lee, Saxe & Branco 2023; [2] Jiang, Xu & Dudman 2022.

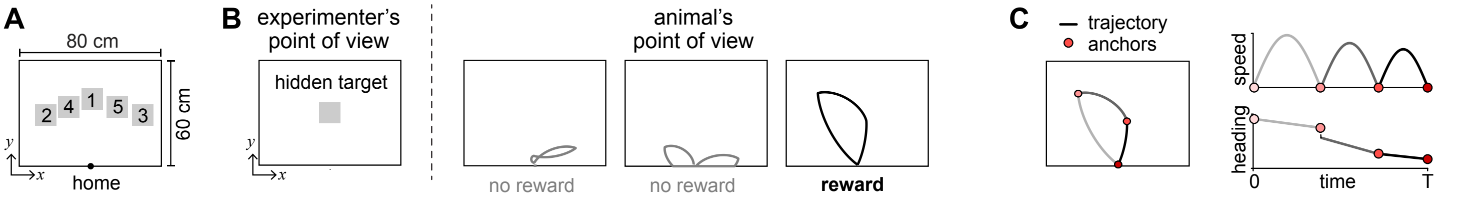

**Figure 1: Task setup.** (A) Mice navigate in an open arena with a hidden target placed at one of five uncued locations (gray boxes). Target locations switch after 5-6 days; switches are also uncued. Upon returning to a home port, mice receive a reward if they intercepted the target. (B) Mice do not receive any localizing feedback about the target location, and must learn based on the outcomes of past trajectories. (C) To mimic animal behavior, we construct a generative model that parametrizes an agent's speed and heading over time (right). The parameters of this model can be chosen to generate a trajectory that links successive pairs of spatial locations ("anchor points", left).

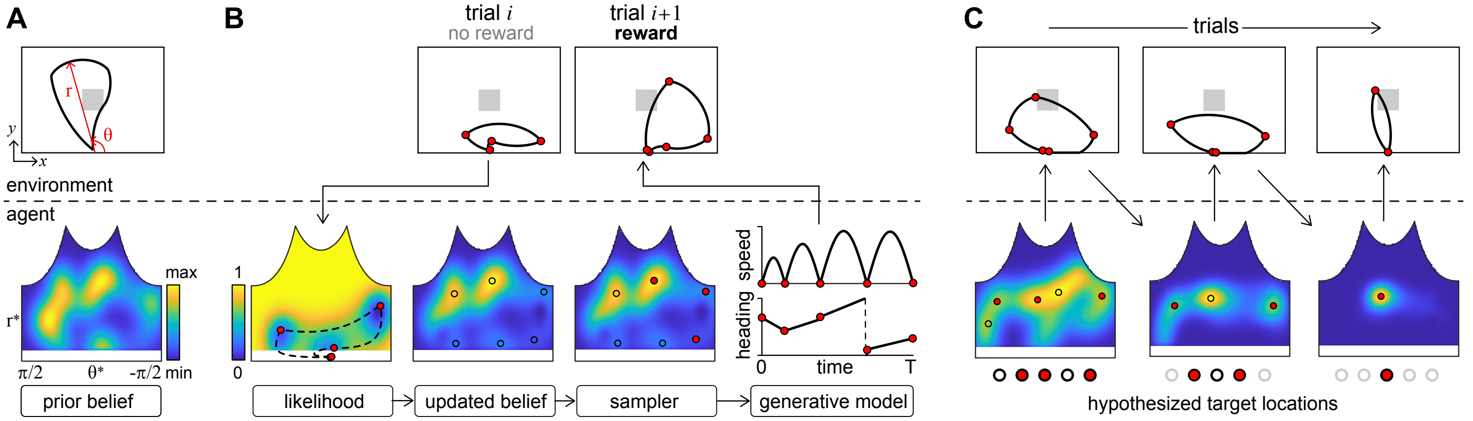

**Figure 2: Agents learn by rapidly narrowing down the space of possible target locations.** (A) We build an agent that maintains a belief about the target location $(r^*, \theta^*)$. (B) Illustration of a single iteration of our algorithm. The agent runs a trajectory on trial $i$ and observes an outcome (no reward). The agent computes the likelihood of observing this outcome given its trajectory (Methods), and uses it to update its belief about the target location (black markers denote local peaks in the posterior belief; note that peaks that were prominent in the prior belief, and that were sampled via the executed trajectory, are nearly eliminated after an unrewarding trial). The sampler selects half of the peaks in the posterior (red markers) to use as anchor points for the subsequent trajectory. The generative model computes the parameters of the speed and heading functions that will link these anchors via smooth paths, which is then used to generate a new trajectory on trial $i{+}1$. (C) The agent rapidly localizes the target by iteratively halving the space of hypothesized target locations. On a given trial, the belief encodes a set of hypothesized target locations (circular markers, lower left). The sampler picks half of these locations (red markers), and uses them to plan the agent's next run (upper left). Upon receiving a reward, the belief update reinforces the sampled peaks that led to reward (lower middle). This continues until the belief localizes on a single target location that is consistent with reward (right column).

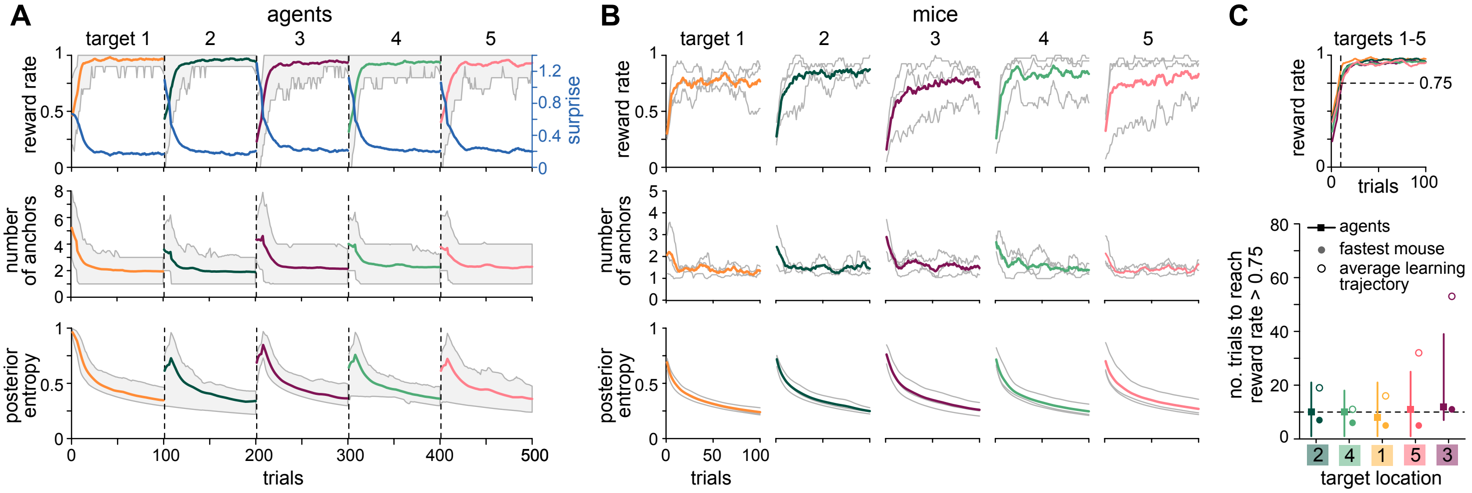

**Figure 3: The speed and temporal evolution of learning mimics mouse behavior.** (A) Agents learn to intercept targets within tens of trials. This is marked by a rapid increase in reward rate (top row), decrease in the number of anchor points that delineate their trajectories (middle row), and decrease in posterior entropy (bottom row, measured relative to a flat prior). Agents use the surprise of new observations under their posterior (blue trace, top row) to detect putative target switches and reset their prior, enabling fast learning of new targets. Individual agent trajectories were smoothed before averaging (window=10 trials); colored lines mark averages over 200 agents; shaded regions mark quantiles of [0.1, 0.9]. (B) Agents qualitatively capture the behavior of individual mice, who refine their trajectories within tens of trials to achieve high reward rates (top row). This is accompanied by a decrease in the number of anchor points (inferred from pauses along their trajectory), and a decrease in posterior entropy (inferred from their executed trajectories and the resulting outcomes). Gray lines show performance of three mice, smoothed over 10 trials and averaged across 4 consecutive days on each target; colored lines show averages across mice. (C) Agents and mice require similar numbers of trials to refine their trajectories across different target locations.