PNAS RESEARCH ARTICLE | NEUROSCIENCE
BIOPHYSICS AND COMPUTATIONAL BIOLOGY

A reinforcement-based mechanism for discontinuous learning

Gautam Reddy[a,b,c,1]

Edited by William Bialek, Princeton University, Princeton, NJ; received September 7, 2022; accepted October 26, 2022

Problem-solving and reasoning involve mental exploration and navigation in sparse relational spaces. A physical analogue is spatial navigation in structured environments such as a network of burrows. Recent experiments with mice navigating a labyrinth show a sharp discontinuity during learning, corresponding to a distinct moment of "sudden insight" when mice figure out long, direct paths to the goal. This discontinuity is seemingly at odds with reinforcement learning (RL), which involves a gradual build-up of a value signal during learning. Here, we show that biologically plausible RL rules combined with persistent exploration generically exhibit discontinuous learning. In tree-like structured environments, positive feedback from learning on behavior generates a "reinforcement wave" with a steep profile. The discontinuity occurs when the wave reaches the starting point. By examining the nonlinear dynamics of reinforcement propagation, we establish a quantitative relationship between the learning rule, the agent's exploration biases, and learning speed. Predictions explain existing data and motivate specific experiments to isolate the phenomenon. Additionally, we characterize the exact learning dynamics of various RL rules for a complex sequential task.

reinforcement learning | physics of behavior | foraging | navigation

Significance

Long-standing reinforcement learning (RL) algorithms incrementally reinforce rewarding actions through accumulated experience. However, past behavioral experiments and recent experiments with mice navigating a complex maze find a sharp discontinuity in learning, akin to an "aha" moment of sudden insight. The learning mechanism that leads to discontinuous learning curves is unclear. We show that the nonlinear dynamics of RL-based learning together with continuous exploration lead to discontinuous learning curves in tree-like structured environments. We develop a quantitative theory which explains the origin and highlights the generality of the phenomenon. The theory explains existing data and provides specific testable predictions.

As we walk the streets of a city, we rapidly figure out paths to new spots after visiting them a few times. For nesting animals, foraging between new locations and their nests in structured environments is an essential aspect of their survival. Rats constantly navigate within a complex underground network of burrows to expand their stores of food (1). Navigating from point A to point B in a structured space requires different strategies compared to a similar task on a flat, open field. In the latter, navigation often involves geometric calculations of distances and angles based on celestial cues, compasses, or landmarks. In a burrow, on the other hand, a rat needs to learn which way to turn at each intersection and benefits from understanding the relationship between places within the network.

The relational structure of mazes offers a well-controlled experimental paradigm to identify biological algorithms for navigating structured environments. Early laboratory experiments on learning algorithms, and animal behavior at large, involved rats navigating a maze (2–7). Rats rapidly learn to navigate to a rewarding location within the maze, which often develops into a habitual action sequence resistant to subsequent changes such as the addition of a shortcut. These experiments and others led to the hypothesis that learning entailed the fixation of stimulus–response relationships due to a reward (6–9). A parallel set of experiments showed that the structure of the maze could be learned during exploration without any significant reward, termed latent learning (10). Latent learning presumably proceeds through the formation of a "cognitive map," which can be flexibly reused when the animal needs to generalize to a novel situation (11–13). This dichotomy between behavioral stereotypy and flexibility is analogous to the modern dichotomy in computational reinforcement learning (RL) between direct and indirect learning, often implemented using model-free and model-based methods, respectively (14–16). However, the specific learning algorithms that animals use to navigate and the circumstances under which one system or the other is employed remain unclear.

Recent developments in deep-learning-based behavioral tracking methods (17–19) allow for following mice in labyrinthine mazes for extended periods of time. In an elegant experiment (20), mice were allowed to navigate (in the dark) an unfamiliar maze structured as a depth-six binary tree (Fig. 1*A*). In each experiment, a mouse moves freely between a cage (marked as home in Fig. 1*A*) and the maze. Markerless pose estimation (17) is used to track its movements continuously over 7 h. Ten of the twenty mice were water-deprived, and a water reward was renewed every 90 s from a port at one end of the maze (marked as a water droplet in Fig. 1*A*). Results recapitulate the aforementioned studies: Mice exhibit rapid learning and eventually execute a quick action sequence from home to the water port. In addition, mice persistently explore the maze with exploration

Author affiliations: [a]Physics & Informatics Laboratories, NTT Research, Inc., Sunnyvale, CA 94085; [b]Center for Brain Science, Harvard University, Cambridge, MA 02138; and [c]NSF-Simons Center for Mathematical and Statistical Analysis of Biology, Harvard University, Cambridge, MA 02138

Author contributions: G.R. designed research, performed research, and wrote the paper.

The author declares no competing interest.

This article is a PNAS Direct Submission.

[1]Email: gautam.nallamala@ntt-research.com.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2215352119/-/DCSupplemental.

Published November 28, 2022.

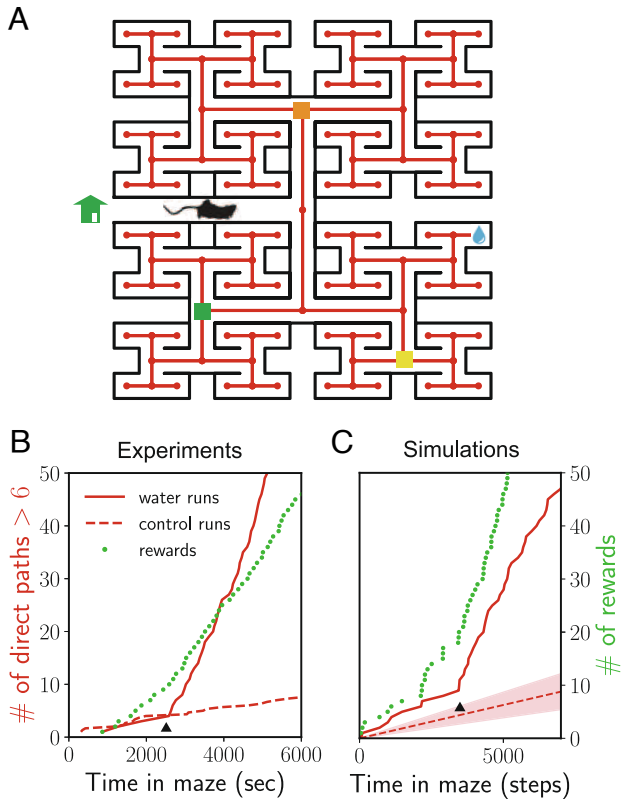

Fig. 1. Discontinuous learning curves in mice experiments and RL simulations. (*A*) A schematic of the depth-6 binary tree maze used in experiments (20) and RL simulations. In each episode of the simulation, the agent begins at home and navigates the directed graph delineated by the maze (red) until it finds the reward. Three intersections (orange, green, and yellow) that the mice have to pass through when executing a direct path of length >6 are marked. (*B*) The cumulative number of direct paths of length >6 (red) and acquired rewards (green) from an individual mouse. The rate of direct paths shows a discontinuity at a distinctive moment (black arrow). The dashed red line corresponds to length >6 direct paths to control nodes. (*C*) Same as in (*B*) for RL simulations. See *SI Appendix*, Fig. S2 for more examples.

biases which are remarkably consistent across rewarded and unrewarded animals.

Intriguingly, the probability that mice take a direct path of >6 correct binary choices toward the water port exhibits a sharp discontinuity, similar to an "aha" moment of sudden insight (Fig. 1*B*), and persists for the rest of the experiment. This moment can occur well after the animal acquires reward for the first time, which distinguishes this phenomenon from one-shot learning. Discontinuous learning curves have also been measured in a variety of other behavioral experiments (21). RL algorithms reinforce correct actions in increments through accumulated experience. This intuition would suggest that RL-based learning is presumably incompatible with step-like learning curves. The availability of the full history of decisions made by mice within the maze presents a unique opportunity to identify the mechanism behind step-like learning curves.

In this manuscript, we use numerics and analytical calculations to rationalize the empirically observed discontinuous learning curves and identify environmental architectures where we should generically expect such discontinuities. We present four main contributions. First, we use inverse RL to decouple and analyze the influence of reward-based learning on the exploratory behavior measured in ref. 20. In this setting, we show using agent-based simulations that persistent exploration combined with simple RL rules reproduce discontinuous learning. Second, we develop a general framework for RL-based sequence learning on tree-structured relational graphs. We use this framework to explain why RL algorithms will generically lead to discontinuous learning curves in such structured environments. Third, we develop a nonlinear, continuous-time model, which accurately captures the dynamics of reinforcement propagation in different exploration regimes. This model extends to commonly used model-free and model-based variants of RL, whose dynamics are analytically quantified. Finally, a reanalysis of experimental data lends further support for the theory and motivates specific experiments to isolate the phenomenon.

Results

Discontinuous Learning in RL Simulations. We begin by specifying an RL model closely following the experimental setup of (20) (Fig. 1*A*). The model is defined by the states (s), how the state changes when a certain action (a) is taken, and the expected reward for each state–action pair, $r(s, a)$. The states determine the information the agent can use to make a decision. Consistent with the history dependence of the exploratory behavior measured in experiments, we assume that the agent knows which specific intersection it is currently at and where it is coming from. That is, the states are the directed edges of the graph that delineates the maze in Fig. 1*A*. When the agent arrives at an intersection along a certain corridor, it has three choices: It can choose to continue along either of the two corridors at that intersection or back where it came from. A fixed reward (r) is delivered in the corridor leading to water.

Upon finding the reward, the agent is reset at the starting point (marked in Fig. 1*A*) and the simulation is repeated. This episodic formulation departs from the experimental setting; we find that an agent placed in an environment with delayed reward renewal (as in the experiment) often learns a degenerate policy which oscillates back and forth at the water port for the rest of the simulation. Of course, a mouse recognizes that water does not immediately reappear after it has been consumed (even if it does not know the precise renewal time) and explores the maze before eventually returning to the water port. For simplicity, we have used an episodic formulation instead of explicitly modeling this time delay.

An RL model is specified by the policy and the learning rule. We use a modified version of the standard softmax policy (14), which chooses actions with a log-probability proportional to their expected long-term reward or value, $q(s, a)$, of taking action a at state s. Specifically, actions are chosen randomly with probability $\pi\ (a|s) \propto e^{q_\varepsilon(s,\ a)+q_r(s,\ a)}$ up to a normalization constant. Here, we have split $q(s, a)$ into two terms, $q_\varepsilon(s, a)$ and $q_r(s, a)$. q_ε is the intrinsic value the agent receives on taking an action at that state and is kept fixed throughout learning. q_r is the extrinsic value, which is initially set to zero and is modulated by reward-based learning. Before learning, the agent makes stochastic exploratory choices based on $q_\varepsilon(s, a)$, which is presumably set by an innate bias or guided by knowledge external to the present task. This term is included in our RL model to explain the observed exploratory behavior of unrewarded and rewarded mice. As learning progresses, these exploratory choices are influenced by the reward, which biases the agent toward rewarding actions (or avoids costly ones). The randomness of the policy is set by the magnitudes of q_ε and q_r, whereas the influence of the reward on exploration is set by their ratio.

This split between intrinsic and extrinsic rewards allows us to examine in silico the influence of a learning rule on natural behavior. We first determine q_ε from the behavior of

unrewarded mice in experiments using maximum entropy inverse RL (MaxEnt IRL 22, 23, *SI Appendix*, Fig. S1*A*). MaxEnt IRL finds the maximum entropy policy and the associated reward function that best explain observed behavioral trajectories (see *Methods* and *SI Appendix* for a brief overview of MaxEnt IRL). Next, we enable learning by specifying a biologically plausible temporal-differences learning rule (14, 24–27). Specifically, q_r is updated using the learning rule:

$$\begin{aligned} &q_r(s,a) \rightarrow q_r(s,a) + \alpha\delta, \quad \text{where} \\ &\delta = r - q_r(s,a), \ \text{at the goal state,} \\ &\delta = \gamma\langle q_r(s',.)\rangle_\pi - q_r(s,a), \ \text{otherwise.} \end{aligned} \qquad [1]$$

δ is the reward prediction error, and the expectation above is with respect to the policy the agent uses at the next state (s'). The discount factor γ, which takes values between 0 and 1, is commonly used to introduce an effective time horizon and regularize the value function. Since our stochastic policy implicitly regularizes the value, we set $\gamma = 1$ throughout this paper. By comparing the best fit q values obtained from MaxEnt IRL for rewarded and unrewarded mice, we estimate the reward as $r \approx 2$ (*SI Appendix*, Fig. S1 *C* and *D*). The remaining free parameter, α, scales the rate of learning. Similar to the learning curves from experiments shown in Fig. 1*B*, we track the cumulative number of rewards acquired by the agent and the cumulative number of long direct paths (length >6) to the goal from distant locations in the maze.

Simulated RL agents exhibit rapid learning similar to those observed in experiments. Importantly, the rate of taking a long direct path deviates discontinuously from the default rate (i.e., as expected from pure exploration) at a distinctive moment during learning, reproducing the "sudden insight" phenomenon observed in experiments (Fig. 1*C*). This phenomenon is reproduced during reruns with variability comparable to the variability observed across mice in experiments (*SI Appendix*, Fig. S2*A*). Fitting the rate of direct paths using a logistic function, we find that the transition can be localized to within fewer than three trials in about half of the runs (*SI Appendix*, Fig. S2*B*).

Goal-Oriented Navigation on Tree-Like Relational Graphs. To identify the mechanism that underpins the sharp transition in learning, we now develop a framework for goal-oriented navigation on tree-like relational graphs. We use this framework to reproduce the discontinuous learning phenomenon, develop a mathematical theory that captures the learning dynamics, and highlight the essential ingredients that lead to the phenomenon.

In this task, the agent traverses a relational graph (a directed graph whose edge labels specify the action or relationship between two states) from a fixed starting point to a goal where it receives a reward (Fig. 2*A*). We track its progress in finding the direct path (highlighted in Fig. 2*A*) by accumulating experience across multiple episodes. We wish to consider graphs that capture the core features of a structured environment such as roads on a university campus or abstract knowledge graphs (28). Specifically, we require 1) discrete decision points and choices; 2) the graph is sparse; namely, the number of paths of comparable length to the direct path is small (unlike a Manhattan-like grid); and 3) long, branching side paths which lead to dead ends.

A large class of graphs that satisfy the above three requirements and are yet sufficiently simple to allow for an in-depth quantitative analysis are tree-structured graphs (Fig. 2*A*), which include the maze architecture from the experiments. Simulating an RL agent in a balanced ternary tree (Fig. 2*B*), we find a sharp discontinuity in the rate of taking the direct path from the start to the goal. Examining the dynamics of reinforcement propagation shows that the reinforcement signal primarily propagates along the direct path (Fig. 2*B*) and that the discontinuity occurs precisely when the reinforcement signal reaches the start (Movie S1). In contrast, an RL agent in a Manhattan-like 6 × 6 grid leads to diffuse propagation of the reinforcement signal and a smooth learning curve (Fig. 2*C* and Movie S2). In *SI Appendix*, Fig. S3 and Movies S3 and S4, we present the learning curves for four additional architectures: a binary tree where the length of the corridors agent is explicitly modeled, a binary tree where the agent is allowed to reverse its direction, and two random graphs with different sparsities. We observe discontinuous learning curves for all of these architectures except for the dense random graph, highlighting that the task structure plays a role in whether discontinuous learning curves are observed.

The structure of tree-like graphs enables us to identify elements of the graph topology and learning dynamics that lead to discontinuous learning. The key insight is that the full complexity of sequence learning on a tree-like graph can be reduced to analyzing the learning dynamics on a simpler linear track with side paths represented as single nodes, as shown in Fig. 3*A*. Specifically, recall that for tree-like graphs, the side paths necessarily lead to dead ends. On encountering a dead end, the agent will turn back and eventually reencounter the direct path. The agent's movements in a side path can thus be represented as a single node noting that if the agent goes in, it will surely return back. When the agent returns back from the side path, it can either choose to go toward or away from the goal.

We emphasize two points that allow this simplification. First, even though we have used a single node with reflecting boundaries to represent the side paths (Fig. 3*A*), an agent may spend a considerable amount of time exploring each of these side paths. Since the time spent within the side path does not influence reinforcement propagation on the direct path, we can safely assume that the agent spends a single step on the side path. Note that the discontinuity in learning is sharper if we suppose that the agent spends longer than a single step in each side path. Second, as long as the side path is sufficiently long, it is unlikely that the reinforcement signal will propagate through the entire side path and bias the agent to go into the side path. Therefore, we may ignore the details of the dynamics within the side path and assume that the q_r value of going into the side path remains at zero. It is important to note that the agent may still learn to turn toward the goal when exiting a side path.

The agent's exploration biases (specified by q_ε) play an important role in determining the qualitative character of the learning dynamics. A key parameter is the probability of continuing toward the goal along the direct path whose corresponding q_ε value we denote ε (Fig. 2*B*). We have assumed a homogeneous ε for simplicity. The discontinuous learning phenomenon is still observed if this assumption is relaxed (see for example Fig. 1*C* where the empirically derived q_ε values are heterogeneous). By varying ε, we examine how the agent's initial exploration and learning dynamics depend on the agent's bias toward taking the correct actions. When $e^{\varepsilon} \gg 1$, the agent continues on the direct path for long stretches and rapidly reaches the goal. In this trivial case, the graph effectively reduces to a linear track without side paths that stretches from the starting point to the goal. In the opposite limit, $e^{-\varepsilon} \gg 1$, correct actions along the direct path are rare. To make progress, the agent would have to take constant detours toward the goal through side paths, whose probability is set by the corresponding value $q_\varepsilon = \varepsilon'$ (Fig. 3*A*).

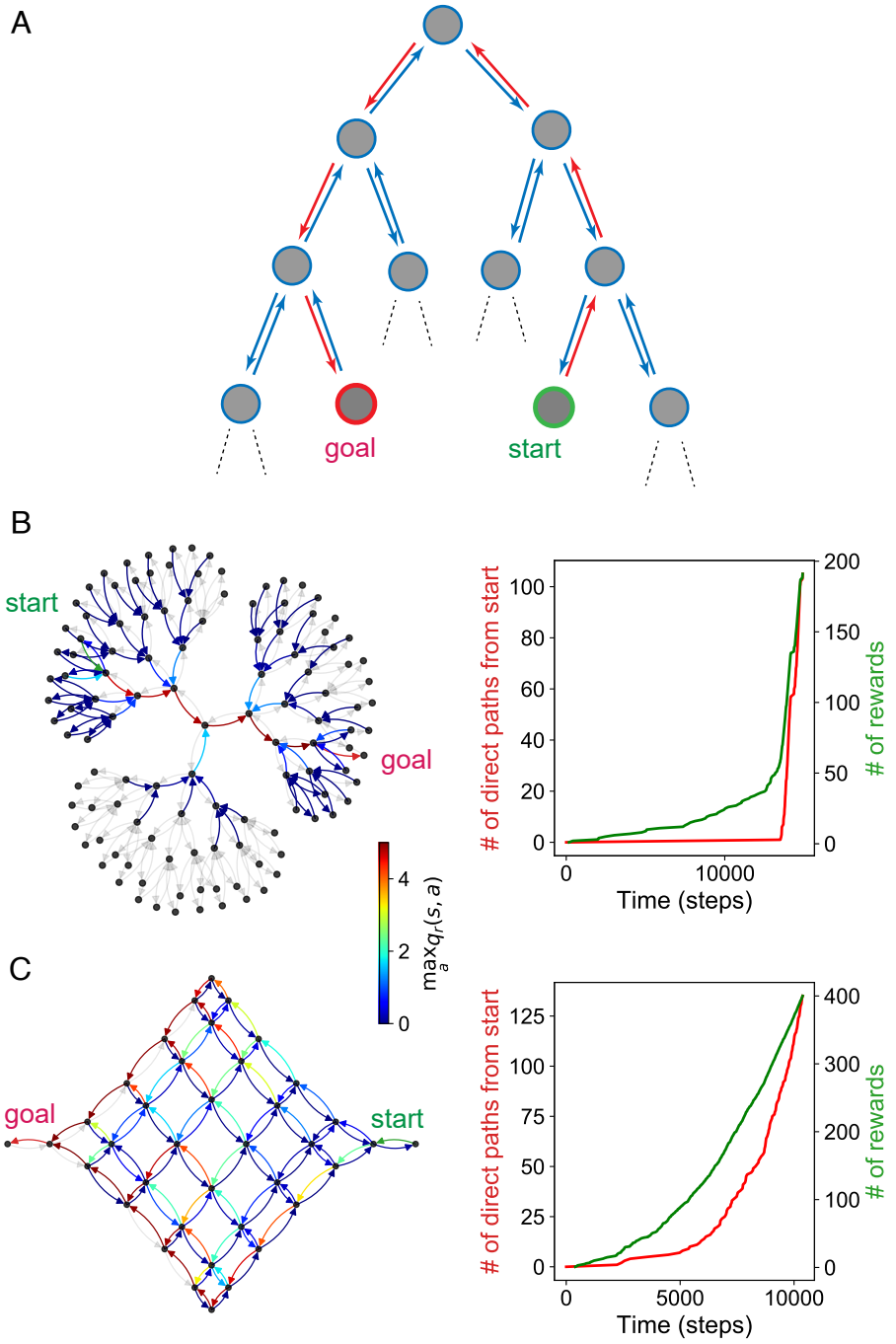

Fig. 2. Reinforcement waves during sequence learning on relational graphs. (*A*) We consider a task where the agent traverses the directed edges of a relational graph to navigate from start (green) to goal (red). The direct path from start to goal is highlighted in red. (*B* and *C*) A discontinuous learning curve for a balanced ternary tree and a smooth learning curve for a Manhattan-like 6×6 grid. In all simulations, we use a standard softmax policy ($q_\epsilon = 0$) with $\alpha = 0.1, r = 5$. The colors show the q_r value of the best action at each state (directed edge). The gray edges have maximum q_r value less than 10^{-3}.

Clearly, if the probability of going toward the goal both along the direct path and through side paths is small ($e^{-\varepsilon'}, e^{-\varepsilon} \gg 1$), the agent is very unlikely to make it to the goal. Thus, whether the agent makes any learning progress whatsoever will depend on the exploration biases. We find that for large graphs, the exploration statistics display three sharply delineated regimes depending on the net probability of going toward the goal vs. back toward the start (*SI Appendix*). If this net probability is negative, the "cautious" agent constantly returns to the starting point and does not learn the task. When the net probability is positive, the "adventurous" agent on average ventures closer to the goal. The marginal case of zero net probability leads to diffusive exploration.

The Mechanistic Basis of Discontinuous Learning Curves. We now examine the learning dynamics generated by the rule Eq. **1**, beginning with RL simulations on the reduced architecture shown in Fig. 3*A* followed by a theoretical analysis. Since actions that lead the agent away from the goal are never reinforced during learning, only the q_r values for continuing along the direct path toward the goal (q_n) and turning toward the goal when exiting the side path (q'_n) at each intersection n should be tracked

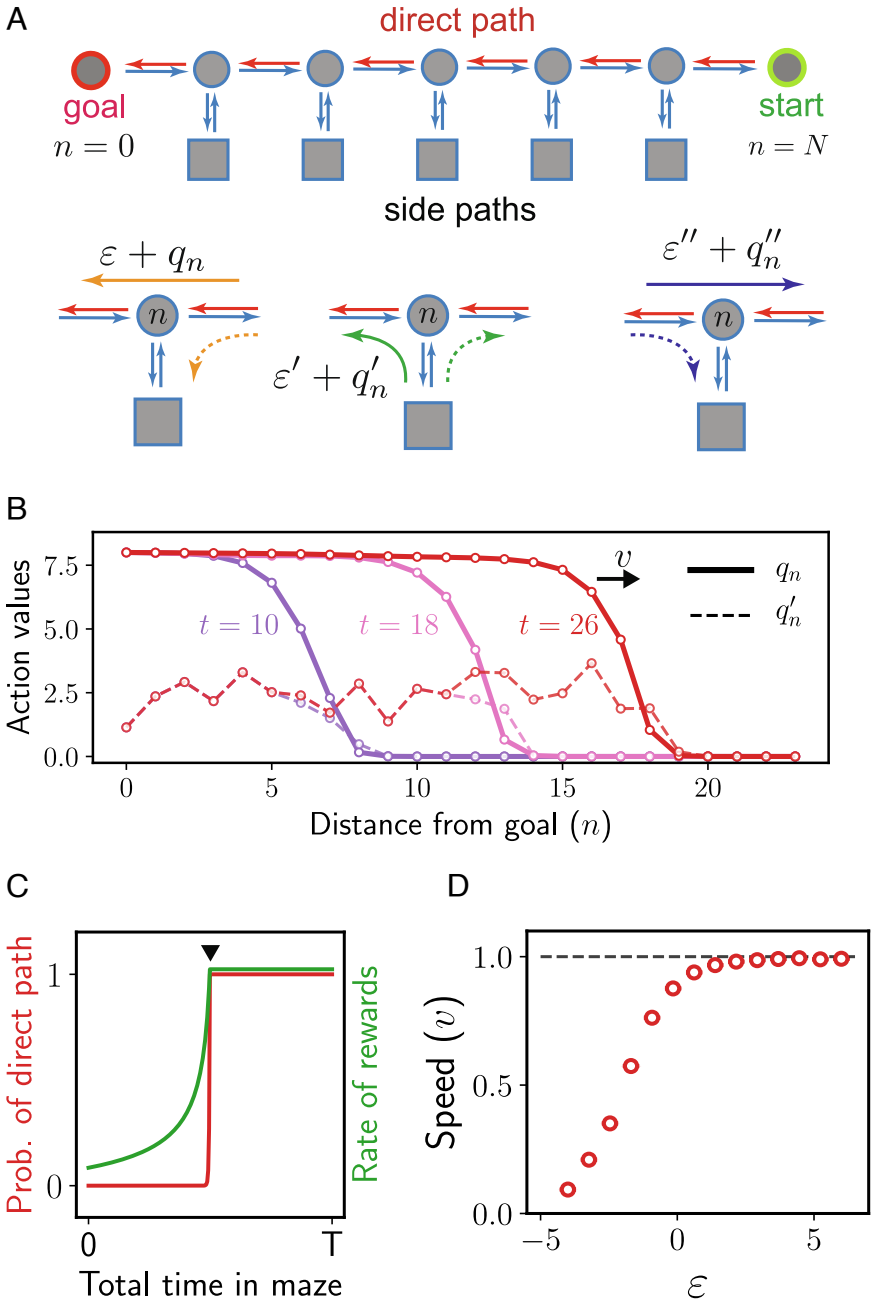

Fig. 3. Wave-like reinforcement propagation during sequence learning on treelike graphs. (*A*) An equivalent representation of the tree-structured graph in Fig. 2*A* highlighting the direct path and the possible branches into side paths at each intersection along the direct path. Note that while each side path is shown as a single node with reflecting boundaries, these represent long detours which will lead to a dead end, forcing the agent to turn back and eventually return to the direct path. The exploration biases $\varepsilon, \varepsilon', \varepsilon''$, and the corresponding reward-modulated biases q_n, q'_n, q''_n for the three cases of going toward the goal on the direct path (*Left*), toward the goal from the side path (*Middle*) and away from goal on the direct path (*Right*) are shown. (*B*) The learned values q_n and q'_n for three snapshots showing the propagation of the reinforcement wave. (*C*) An illustration showing the discontinuity in the probability of a direct path and the rate of rewards. The discontinuity occurs the moment the wave hits the starting point (Movie S5). (*D*) The speed of the wave for a range of ε. Smaller ε values correspond to more difficult tasks.

(We use $n = 0$ and $n = N$ for the goal and start respectively, see Fig. 3*A*). Fig. 3*B* shows q_n and q'_n at three time points (in units of $1/\alpha$ episodes), highlighting the wave-like propagation of the value, q_n (Movie S5). The learning curves show a sharp discontinuity (Movie S5 and Fig. 3*C*), which occurs precisely when this wave reaches the starting point. Total learning time is determined by the wave's speed, which we measure as the number of intersections on the direct path the wave crosses every $1/\alpha$ episode. Tracking the half-maximum of q_n, we find that the wave travels at a constant speed, v (Movie S5). Simulations across a range of ε show the speed saturating at $v = 1$ for $\varepsilon \gtrsim 1$, which decreases to zero with decreasing ε (Fig. 3*D*), hinting at distinct regimes. The factors that determine the speed and profile of the wave will be discussed in the following section.

The origin of discontinuous learning and "reinforcement waves" can be intuitively understood by examining how learning operates at each intersection. We highlight three factors: 1) The correct action at an intersection is reinforced only if the action at the subsequent intersection is reinforced, implying that the chain of reinforcement has to travel backward from the goal; 2) When

an intersection is sufficiently reinforced, the probability of the correct action at that intersection increases by a large factor as long as the reward is sufficiently large ($e^{r+\varepsilon} \gg 1$). Since the rate of traveling directly from start to goal is the product of the probabilities of taking the correct action at each intersection, this rate will increase rapidly when the wave reaches the start, and 3) if the agent is unlikely to take the correct action at a certain intersection ($e^{-q_\varepsilon} \gg 1$ for that action), reinforcement is applied through a few rare events until the intrinsic bias is overcome, $q_r + q_\varepsilon > 0$. Since the probability of taking the correct action in turn increases rapidly with reinforcement, the learning curve for taking the correct action at each intersection will appear step-like.

The first factor emphasizes why we should expect the reinforcement signal to propagate backward from the goal to the starting point. The second factor highlights the fact that the observable (i.e., the probability of taking the direct path) is a steep, nonlinear function of the underlying dynamical variables. The third point explains why the wave front has a steep profile (Fig. 3*C*). Put together, these three factors imply that when the task is nontrivial, the wave of reinforcement marches backward from the goal, reinforcing correct actions, one intersection at a time with step-like learning at each intersection. The observed discontinuous transition in learning occurs when the wave reaches the starting point.

A Nonlinear, Continuous-Time Model Accurately Captures the Dynamics of Reinforcement Propagation. This intuitive picture can be made mathematically precise by examining the effects of the learning rule, Eq. **1**, on q_n and q'_n. We summarize the results here; refer *SI Appendix* for full details. When $\alpha \ll 1$, we find that their expected change, $\dot{q}_n, \dot{q}'_n$, over $1/\alpha$ episodes is given by

$$\begin{aligned} \dot{q}_n &= \mu_n \left(\sigma_{n-1} q_{n-1} - q_n\right), \\ \dot{q}'_n &= \mu'_n \left(\sigma_{n-1} q_{n-1} - q'_n\right), \end{aligned} \qquad [2]$$

where μ_n and μ'_n are the average number of times per episode the agent crosses intersection n through the direct path or the side path, respectively, and σ_n is the probability of continuing along the direct path at intersection n. In general, μ_n and μ'_n depend on the transition probabilities and thus the values at every intersection in the graph. The analysis is made tractable by noticing, first, that the ratio μ_n/μ'_n is determined by the relative probability of taking the correct action at intersection n through the direct path vs. the side path. Second, no learning occurs outside of the front and bulk of the wave. Finally, learning at the front of the wave happens only when subsequent intersections are already sufficiently reinforced, which implies that the agent is likely to go directly to the goal immediately after crossing the front. Thus, in each episode, the intersection at the wave's front is crossed just once on average, $\mu_n + \mu'_n \simeq 1$. This relation combined with the expression for μ_n/μ'_n fixes μ_n, μ'_n. The q_n, q'_n's obtained from numerical integration of Eq. **2** are in excellent agreement with the ones from full-scale RL simulations (Fig. 4*A*). An analysis of Eq. **2** reveals two qualitatively distinct regimes of wave propagation with $e^{\varepsilon} \gg 1$ and $e^{-\varepsilon} \gg 1$ as their asymptotic limits. We term these the expanding and marching regimes, respectively. Maze architectures that could exhibit these two regimes are illustrated in Fig. 4 *B* and *C*.

The expanding regime ($e^{\varepsilon} \gg 1$) corresponds to the trivial case where the agent is likely to traverse straight from the starting point to the goal. Eq. **2** leads to linear dynamics in this regime, which can be solved exactly. We find $q_n(t) = rP(n, t)$, where $P(n, t)$ is the regularized lower incomplete gamma function. For large n, the half-maximum is at $n_{1/2} = t$, which explains the speed $v = 1$ observed in simulations for $\varepsilon \gtrsim 1$, and the width of the profile expands with time as $\sqrt{t}$.

In the marching regime ($e^{-\varepsilon} \gg 1$), the negative ε leads to qualitatively different, nontrivial dynamics. Any step on the direct path that has previously been reinforced beyond $|\varepsilon|$ is more likely to be traversed. When the reinforcement wave reaches an intersection p on the direct path that is yet to be reinforced to $|\varepsilon|$, the reinforcement of that step occurs through rare events until $q_p \simeq |\varepsilon|$. Meanwhile, the direct path for $n < p$ is rapidly reinforced. The rare events at p combined with rapid reinforcement for $n < p$ lead to a bottleneck at p and a steep wave profile. Once q_p reaches $|\varepsilon|$, it is subsequently reinforced rapidly and q_{p+1} in turn begins to be slowly reinforced through rare events. Thus, the wave "marches" forward reinforcing one step at a time. Computing the duration τ it takes to march one step will let us estimate the speed of the wave, $v = \tau^{-1}$.

The duration τ can be calculated by examining the nonlinear dynamics in the front ($n = p$) and bulk ($n < p$) of the wave (*SI Appendix*). The full dynamics in the bulk play a role as the reinforcement received at the intersection $n = p$ depends on the temporal dynamics of q_{p-1}, which in turn depends on q_{p-2}, and so on. However, it can be shown that the dynamics in the bulk are linear and exhibit self-similarity with period τ. Exploiting a conservation equation that results from these properties, we compute the wave speed as

$$v = \tau^{-1} = \frac{r}{r + e^{|\varepsilon|} - 1}, \qquad [3]$$

which is in excellent agreement with the speed measured in RL simulations (Fig. 4*D*). The wave profile in the bulk is given by $q_{n-1}(t) = r - \beta(r - q_n(t))$, where $\beta = -\tau^{-1} W(-\tau e^{-\tau})$ and $W(x)$ is the Lambert W function. Most of the learning at a certain intersection occurs in $\lesssim 1/\alpha$ episodes (Fig. 4*E*). Since the wave speed is less than one in the marching regime, each intersection is almost fully reinforced before the wave marches to the next one, thus quantifying the aforementioned intuitive argument that a step-like learning curve is observed at each intersection.

The results are summarized in Fig. 4*F*, which depicts the expanding and marching regimes in addition to the "stalled" regime corresponding to the exploration parameters where learning is largely absent.

Other Learning Rules Lead to Reinforcement Waves with Altered Speeds and Profiles. Common variants of the *SARSA* rule (14) in Eq. **1** also lead to discontinuous learning via reinforcement waves, highlighting the generality of the phenomenon. A detailed analysis of each of these variants is presented in *SI Appendix*, which we summarize here.

We find that Watkins' Q-learning, which uses a slightly modified version of the rule Eq. **1**, leads to largely similar wave speeds and profiles. The advantage of Q-learning is that the q_r values can be learned off-policy, i.e., the agent's behavior is not necessarily derived from the learned q_r values. To decouple the influence of learning on behavior, we use Q-learning together with an explorative agent that disregards the learned q_r values. We find expanding waves irrespective of the exploration bias, suggesting that expanding waves are the "default" dynamics without feedback in the structured environments considered here. Feedback due to learning leads to traveling waves with steeper profiles as observed in the marching regime. Both Q-learning and *SARSA* learn values from local updates, which constrains the wave speed to be at most one.

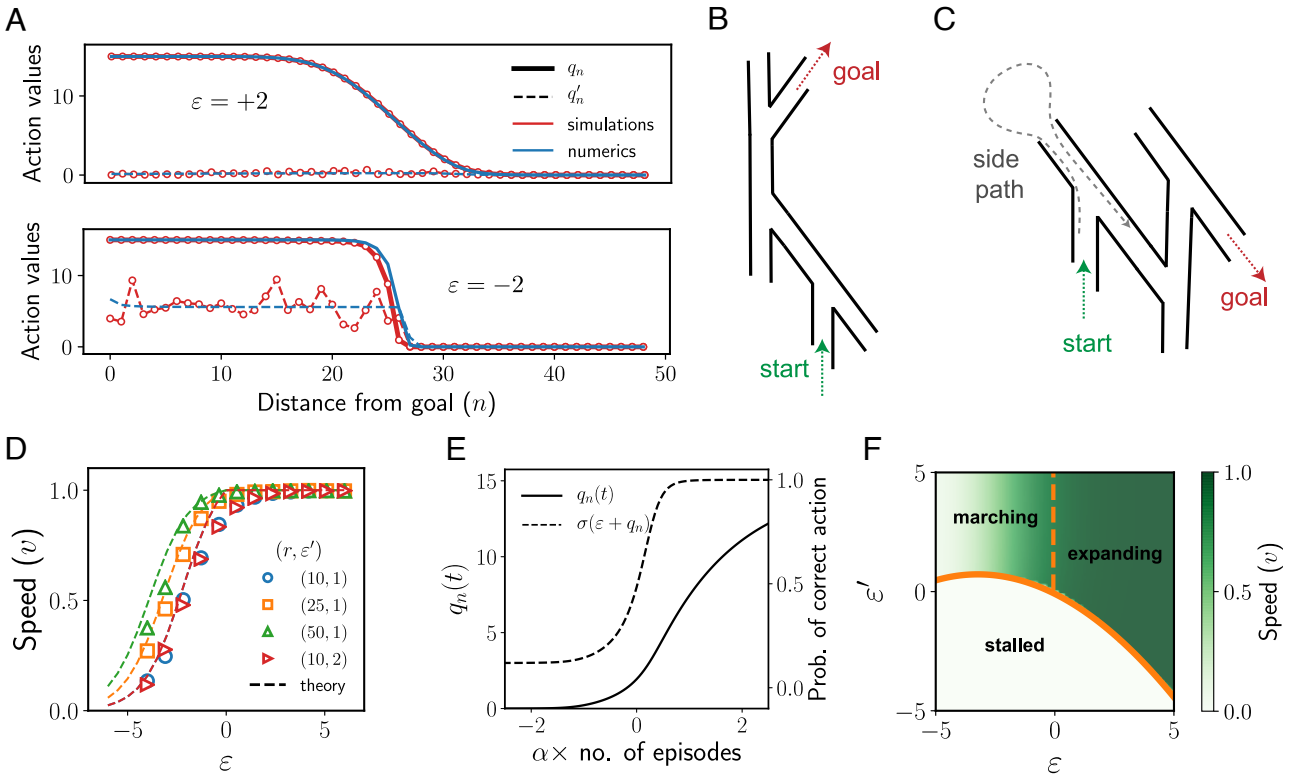

Fig. 4. The expanding and marching regimes of wave propagation: numerics and theory (*A*) A snapshot of q_n and q'_n for $\varepsilon = \pm 2$, shown in red and blue from RL simulations and from numerically integrating Eq. **2**, respectively. (*B* and *C*) Illustrations of how mazes with $e^{\varepsilon} \gg 1$ (*B*) and $e^{-\varepsilon} \gg 1$ (*C*) can be constructed. (*D*) The theoretical prediction for the speed (dashed lines) closely aligns with the speed measured in the full RL simulations. The red and blue dashed lines are aligned. (*E*) The change in $q_n(t)$ when the wave passes through intersection n, shown here for $\varepsilon = -2$. The x-axis is centered at the moment when the probability of taking the correct action, $\sigma(\varepsilon + q_n) = 1/2$. Note that learning at this intersection is localized to $\lesssim 1/\alpha$ episodes. (*F*) The distinct learning regimes for a range of exploration parameters. Here, $N = 20, \varepsilon'' = 0$. We use $r = 15$ for panels *A* and *E* and $r = 12$ for panel *F*.

An alternative class of models build a model of the environment from experience, similar to a cognitive map, and update the values offline by sampling from the model (planning). We consider *Dyna-Q*, which implements a simple version of this general idea. Specifically, *Dyna-Q* first learns a model of future states and rewards for every state–action pair it encounters during the task. At each step, it samples n_p state–action–state–reward transitions from the model and updates their corresponding values. We show that *Dyna-Q* applied to our setting leads to the same behavior as Eq. **1** with an enhanced learning rate $(1 + n_p)\alpha$. Intuitively, when the agent plans, learning, which otherwise occurs only through physical exploration, is sped up due to mental exploration. However, since both physical and mental exploration employ the same search process, the result is a simple scaling of the learning rate.

Another common variant with nonlocal updates is *SARSA* combined with eligibility traces, which are an efficient, biologically plausible mechanism for enhancing learning speed when rewards are sparse (14, 29). Instead of updating the value of the current state–action pair, eligibility traces effectively use the current reward prediction error to also update the k most recent state–action pairs. The exact learning dynamics can be calculated (*SI Appendix*) and are qualitatively similar to the *SARSA* case. In the expanding regime, eligibility traces scale the wave speed by a factor $1 + k$. The speed in the marching regime has a nontrivial, sublinear relationship with k (*SI Appendix*, Fig. S4*B*), which can be computed from the theory using a self-consistent equation (*SI Appendix*). Intuitively, the speed increases with k since the front of the wave receives reinforcement from the intersection $1 + k$ steps along the direct path, which has a larger value compared to the subsequent intersection. In the limit $k \to \infty$, we show that the speed converges to a maximum $v_\infty = r/(|\varepsilon| + e^{|\varepsilon|} - 1)$.

The theoretical predictions for the various learning rules are verified in simulations (*SI Appendix*, Figs. S4 and S5).

Experimental Tests

In addition to reproducing the discontinuous learning curves observed in experiments, the theory provides predictions which can be immediately tested by reanalyzing the data from ref. 20. Specifically, note that the learning curves in Fig. 1*B* correspond to the number of direct paths greater than a certain length, namely, six. If the discontinuity in the learning curves is due to a reinforcement wave, this discontinuity should occur at a later time for direct paths beginning from farther nodes. This prediction should be contrasted with an alternative mechanism where sudden insight corresponds to the singular moment when the mouse has figured out the global structure of the environment and uses this knowledge to find direct paths from distant sections of the maze. The experimental data lend support for the former hypothesis, which shows that the discontinuity is delayed for longer direct paths (Fig. 5*A*). The time delay between these discontinuities provides an estimate of the wave speed. The smaller rate of taking direct paths for longer paths observed in Fig. 5*A* can also be explained in our framework. The reward (estimated as $r \approx 2$ previously) is not sufficiently large to fully overcome the stochastic, exploratory drive of the agent, leading to

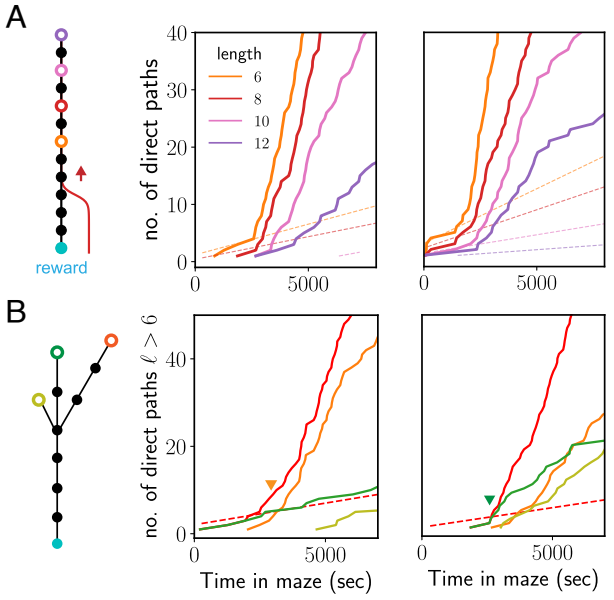

Fig. 5. Wave propagation is consistent with experimental data. (*A*) Theory predicts that the reinforcement wave reaches locations further away from the goal at later times. Shown here is the cumulative number of direct paths of lengths at least 6, 8, 10, and 12 in orange, red, pink, and purple, respectively for two mice. The dashed lines are direct paths to control nodes. (*B*) Stochasticity in exploration and learning dynamics can lead to the wave reaching different intersections at different moments during learning. Shown here are direct paths of length >6 from three distinct intersections in the maze (marked in Fig. 1*A* with their respective colors) for two mice.

a significantly smaller probability of taking a longer direct path. This decreasing probability provides an estimate for the range of wave propagation, N_{range}. The theory predicts that N_{range} and the speed of wave propagation should increase with increasing reward for $e^{-\varepsilon} \gg 1$, which can be tested in future experiments. An intriguing possibility is to observe the transition in speed from the expanding to marching regimes by manipulating the exploration biases, for example, by modifying the inclinations of the T-junctions in a complex maze (as illustrated in Fig. 4 *B* and *C*) or manipulating the number of branches at each intersection.

A potentially important confounding factor for observing a single, distinct discontinuity in the learning curves is when multiple paths of length comparable to the direct path are available. The speed at which the wave propagates along these competing paths depends on a number of factors, including their number, lengths, and the exploration statistics within each path. If a competing path is fully reinforced earlier than the direct path, it can interfere with learning the direct path. Multiple paths can explain the variability observed in experimental trajectories. Indeed, the learning curves in Figs. 1*B* and 2*C* effectively average over all direct paths of certain lengths. If paths of similar lengths from distant nodes exhibit discontinuities with only slight delays, the averaged curve will appear smoother than when each path is observed separately. Consistent with this intuition, considering paths from specific locations in the experiment highlights the variability across mice in which of these paths contributes most to the discontinuity (Fig. 5*B*).

Additional experiments designed similar to our setting in Fig. 2*A* will provide crucial data to resolve sources of variability. Specifically, our analysis suggests examining direct paths between two specific start and goal locations in an episodic setting or equivalent. This will ensure that the measured learning curves do not reflect contributions from different locations in the maze and highlight the passage of the wave along the direct path between these two nodes. Further, learning via reinforcement is not necessarily monotonic in the experimental setup of ref. 20, which makes it challenging to infer the progression of learning at each intersection directly from data. For example, if the animal samples the water port when reward is absent, the resulting reinforcement can be negative, which leads to unlearning of the path toward the reward. This nonmonotonicity is absent in an episodic setting and will lead to a clearer interpretation of the learning curves at each intersection.

Discussion

The discontinuous learning phenomenon observed in complex mazes and other learning tasks clashes with the intuition that RL-based algorithms make learning progress by incrementally reinforcing rewarding actions. Here, we have shown that a standard biologically plausible RL rule consistently reproduces this phenomenon in simulations designed to reflect maze experiments and more generally during goal-oriented navigation in large, tree-like relational graphs. In such environments, the value signal propagates as a steep, traveling reinforcement wave, which sequentially reinforces correct actions along the path toward the goal. Sudden insight occurs the moment the wave reinforces all the correct actions along the main path. Discontinuous learning curves arise due to a combination of the effectively one-dimensional task structure in tree-like structured environments, the local propagation of reinforcement, and the positive feedback of reinforcement on behavior. These factors together with the agent's innate exploration biases determine the dynamics of wave-like reinforcement propagation, including its speed and profile. The exploration biases play an important role as they determine whether any learning occurs in the first place (the stalled regime), and, if learning does progress, whether the learning dynamics are limited by the learning rule (expanding regime) or due to the low probability of taking the correct action (marching regime). While common model-free and model-based variants of the RL rule may enhance the learning speed and alter the wave's profile, the qualitative characteristics of wave propagation are preserved.

Whether and under what contexts animals learn correct actions directly from experience or indirectly through a learned model of the environment is a long-standing debate. The aha moment observed in the experiments of ref. 20 would naively appear to support the latter hypothesis. We have shown here that existing experimental data are consistent with the propagation of a reinforcement wave (Fig. 5), and thus, RL-based direct learning cannot be ruled out. Further experiments should reveal and verify the generality of the discontinuous learning phenomenon. The framework presented in this manuscript should help guide specific experiments to delineate direct and indirect learning (see *Experimental Tests* for further discussion).

We emphasize that the backward propagation of reinforcement arises as a straightforward consequence of the local RL rule applied in an environment where the goal state is the sole source of reward. However, as illustrated in Fig. 2*C* and *SI Appendix*, Fig. S3*D*, not all graph architectures will display discontinuous learning under these RL rules. We have shown that the topology of large tree-like mazes (with appropriate exploration biases) supports discontinuous learning, but we expect to observe the phenomenon more generally if the graph satisfies certain notions of "sparsity" (*SI Appendix*, Fig. S3*C*). This is because the sharp transition in learning is most salient in highly complex mazes

where the direct path is nontrivial and paths other than the direct path are present but are poor solutions.

Competing paths lead to additional complexity, analogous to when a multitude of local minima compete with the global solution in nonconvex optimization problems. Easily accessible competing paths which are of comparable length to the direct path may lead to nontrivial exclusion effects, effectively average out the learning curves, and amplify variability due to minor differences in exploration biases across animals. Sudden, delayed improvements in generalization performance have been recently observed when neural networks are trained to solve small algorithmic tasks, a phenomenon that has been termed "grokking" (30). Preliminary theoretical work (31) suggests that the task structure imposes highly specific constraints on the representations that can achieve perfect generalization, and "sudden insight" occurs when these constraints are fulfilled. This work and ours suggest that nontrivial constraints on good solutions imposed due to task structure might play an important role in the emergence of sudden learning phenomena.

Our analysis provides a complete characterization of the learning dynamics of various RL rules for a nontrivial sequential decision-making task, which is currently lacking. A key challenge in the theoretical analysis of RL algorithms is the feedback of learning on behavior, which makes the data distribution inherently nonstationary. In our setting, the nonstationarity is reflected by the dynamics of the wave during learning. We have shown that the front of the wave effectively acts as an absorbing boundary, which simplifies the analysis considerably. The learning speed is determined by the number of times the learning rule updates the value at the nose of the wave. Since this number itself depends on the value at the nose, the dynamics are nonlinear. In turn, since the value of the subsequent action depends on the value of the later actions within the bulk, the full interactions between the nose and the bulk of the wave will influence learning speed. We show that the learning speed cannot exceed a certain value due to the locality of the learning rule. Relaxing the locality constraint using eligibility traces enhances the learning speed by widening the value differential between the unreinforced action and the distal action from which it receives reinforcement. A model-based method which uses planning scales up the speed simply by scaling up the number of times it updates each action rather than due to a qualitative change in how reinforcement is propagated.

In specifying our model, we have made certain simplifications that do not capture the full complexity of animal learning. First, we have considered a discretized model of the state and action spaces. While this is a standard approximation, animals use continuous spatial representations and motor control. Standard computational RL rules, such as Eq. **1**, have been fruitfully extended to deal with continuous state and action spaces, for instance, using function approximation and policy gradient methods (14). Biologically plausible variants of these extensions have been proposed (32), including for mental exploration in simple mazes (33). We expect the core intuition behind discontinuous learning to hold even for a more realistic model with continuous state and action spaces. A detailed analysis of RL dynamics for a model which takes these various factors into account is beyond the scope of current work (see *SI Appendix*, Fig. S3*A* and Movie S3 for preliminary results). Second, we have assumed that the animal has a unique representation of each corridor in the maze from the outset. Of course, this representation would have to be learned before the animal can assign and update the value of taking different actions at each corridor (34). Our results should still apply if the timescale for "mapping" the environment is faster than the timescale for RL. The hierarchical structure of neural network-based function approximators enables simultaneous learning of representations and values (35), but the timescales on which these processes operate in animals are unknown and presumably much shorter. An exciting future direction is to extend our framework to spatial navigation tasks with other graph topologies or when learning of proper actions is intertwined with the learning of continuous state representations.

Materials and Methods

Extracting Exploration Statistics from Data and Hyperparameters for RL Simulations. We use MaxEnt IRL (see *SI Appendix* for a brief introduction) to infer the exploration biases of unrewarded mice. As discussed in the *Results* section, the state space was chosen as the directed edges of the graph that delineate the maze in experiments, where the root of the tree corresponds to "home." We pooled trajectories from all unrewarded mice, set $\gamma = 0.8$, and split the trajectories to length $T = 12$ (T should be at least the effective horizon $\sim (1-\gamma)^{-1} = 5$ and choosing a large T slows inference). The choice of γ was motivated by the analysis in ref. 20, which showed that a variable length Markovian model typically chooses $\lesssim 5$ previous states to predict mice behavior. The q_ε values are obtained from maximum likelihood estimation, specifically, from $\log p_{\lambda,0}(s,a)$ after optimizing for λ (*SI Appendix*). Note that due to normalization, the q_ε values are determined only up to a constant additive term for each state.

To estimate r, we apply the above procedure to both unrewarded and rewarded mice. We calculate the difference between rewarded and unrewarded animals in the differences of the correct action's q value and the effective q values of the other two incorrect ones [note $q_{\text{eff}}(s,\boldsymbol{\mathcal{A}}) = \log\left(\sum_{a\in\boldsymbol{\mathcal{A}}} e^{q(s,a)}\right)$]. A subset of these values is shown in *SI Appendix*, Fig. S1, which shows that the correct actions leading to the reward have a value differential of ≈ 2. Since the values of actions close to the reward after learning saturate at r, the value differential is an estimate of the reward, $r \approx 2$. To ensure that this estimate is not significantly influenced by the habitual paths that go directly from home to goal, we repeat the above procedure excluding these paths (*SI Appendix*, Fig. S1). The estimate decreases slightly to $r \approx 1.5$. In the RL simulations of the depth-6 binary tree maze, we use $r = 2$ and $\alpha = 0.33$.

Setup and Notation for the RL Framework for Navigation on Tree-Structured Graphs. A tree-structured graph can be cast as a linear track with side paths, as argued in the *Results* section and illustrated in Figs. 2*A* and 3*A*. The linear track consists of $N-1$ nodes on the direct path, $n = 1, 2, \ldots, N-1$. The agent starts each episode at node $n = N$, and the reward is at the goal node $n = 0$. In addition to these nodes, the nodes from $n = 1$ to $N-1$ each have a side path, which we label as $1_b, 2_b, \ldots, (N-1)_b$. The state space of the Markov decision process is the set of directed edges that connect the various nodes and the side paths as shown in Fig. 2*B*. In other words, both the agent's location in the graph and the direction in which it is headed matter. We denote (n_1, n_2) as the directed edge from n_1 to n_2.

The transition dynamics $P(s'|s,a)$ are deterministic (Note, however, that the policy $\pi(a|s)$ is stochastic). At each directed edge, the agent can choose to go along the directed edges emanating from its current node, except for turning back, e.g., the transition $(n+1, n) \to (n, n+1)$ is disallowed. This simplifying assumption does not affect the results as the agent can effectively turn back by going into a side path and returning $(n+1, n) \to (n, n_b) \to (n_b, n) \to (n, n+1)$. The episode begins with the agent at the directed edge $(N, N-1)$. The directed edge pointing toward the goal node, $(1, 0)$, is an absorbing state, i.e., the agent receives a reward r and the episode ends once the agent traverses that edge. We impose reflecting conditions at edges going into the side paths (n, n_b) and the start node $(N-1, N)$.

The agent receives identical intrinsic exploration rewards at every intersection on the direct path. There are three directed edges leading to any node n, and we

thus consider three cases at each node. These three cases are shown pictorially in Fig. 2*B*. Since the agent can take two actions at each step and the policy depends only on differences of q values, we specify the q values for only one of the actions. Note that q values for both actions are taken into account in the agent-based RL simulations throughout the paper. For the purposes of the theoretical analysis discussed in the *SI Appendix*, it suffices to track the q value for only one of the actions as the q value for the other action is almost never reinforced. The notation used for the three cases is introduced (see also Fig. 2*B*).

1. the agent is on the direct path and going toward the goal, $(n+1, n)$: for the action corresponding to the agent continuing toward the goal $(n+1, n) \to (n, n-1)$, we denote $q_\varepsilon \equiv \varepsilon$, $q_r \equiv q_n$.
2. the agent is on the side path n_b and going toward n, (n_b, n): for the action corresponding to the agent turning toward the goal $(n_b, n) \to (n, n-1)$, we denote $q_\varepsilon \equiv \varepsilon'$, $q_r \equiv q'_n$.
3. the agent is on the direct path and going toward the start, $(n-1, n)$: for the action corresponding to the agent continuing toward the start $(n-1, n) \to (n, n+1)$, we denote $q_\varepsilon \equiv \varepsilon''$, $q_r \equiv q''_n$.

The probabilities of taking the action described in each of three cases is denoted $\sigma_n \equiv \sigma(\varepsilon + q_n)$, $\sigma'_n \equiv \sigma(\varepsilon' + q'_n)$, $\sigma''_n \equiv \sigma(\varepsilon'' + q''_n)$, where $\sigma(x) = 1/(1+e^{-x})$ is the logistic function.

Data, Materials, and Software Availability. Previously published data were used for this work (https://doi.org/10.7554/eLife.66175).

ACKNOWLEDGMENTS. G.R. thanks Andrew Murray and Venkatesh Murthy for useful comments on the manuscript. G.R. was partially supported by the NSF-Simons Center for Mathematical & Statistical Analysis of Biology at Harvard (award number #1764269) and the Harvard Quantitative Biology Initiative.

1. J. B. Calhoun, *The ecology and sociology of the Norway rat. Number 1008*. US Department of Health, Education, and Welfare, Public Health Service (1963).
2. M. S. Small, An experimental study of the mental processes of the rat. II. *Am. J. Psychol.* **11**, 133–165 (1900).
3. W. S. Small, Experimental study of the mental processes of the rat. *Am. J. Psychol.* **12**, 206–239 (1901).
4. D. S. Olton, Mazes, maps, and memory. *Am. Psychol.* **34**, 583 (1979).
5. E. C. Tolman, C. H. Honzik, "Insight" in rats. *University of California Publications in Psychology* (1930).
6. H. C. Gilhousen, An investigation of "insight" in rats. *Science* **73**, 711–712 (1931).
7. I. Krechevsky, "Hypotheses" in rats. *Psychol. Rev.* **39**, 516 (1932).
8. C. Leonard Hull, *Principles of Behavior: An Introduction to Behavior Theory* (Appleton-Century, 1943).
9. E. L. Thorndike, *The Fundamentals of Learning* (Teachers College Bureau of Publications, 1932).
10. D. Thistlethwaite, A critical review of latent learning and related experiments. *Psychol. Bull.* **48**, 97 (1951).
11. E. C. Tolman, Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189 (1948).
12. T. E. J. Behrens *et al*., What is a cognitive map?. *Organizing knowledge for flexible behavior. Neuron* **100**, 490–509 (2018).
13. D. Schiller *et al*., Memory and space: Towards an understanding of the cognitive map. *J. Neurosci.* **35**, 13904–13911 (2015).
14. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 2018).
15. D. Bertsekas, *Dynamic Programming and Optimal Control: Volume I* (Athena Scientific, 2012), vol. 1.
16. S. J. Gershman, A. B. Markman, R. A. Otto, Retrospective revaluation in sequential decision making: A tale of two systems. *J. Exp. Psychol.: General* **143**, 182 (2014).
17. A. Mathis *et al*., Deeplabcut: Markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
18. T. D. Pereira *et al*., Fast animal pose estimation using deep neural networks. *Nat. Methods* **16**, 117–125 (2019).
19. J. M. Graving *et al*., DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife* **8**, e47994 (2019).
20. M. Rosenberg, T. Zhang, P. Perona, M. Meister, Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration. *Elife* **10**, e66175 (2021).
21. C. R. Gallistel, S. Fairhurst, P. Balsam, The learning curve: Implications of a quantitative analysis. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13124–13131 (2004).
22. B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, "Maximum entropy inverse reinforcement learning" in *AAAI* (Chicago, IL, USA, 2008), vol. 8, pp. 1433–1438.
23. S. Levine, Reinforcement learning and control as probabilistic inference: Tutorial and review. arXiv preprint arXiv:1805.00909 (2018).
24. Y. Niv, M. O. Duff, P. Dayan, Dopamine, uncertainty and TD learning. *Behav. Brain Func.* **1**, 1–9 (2005).
25. R. P. N. Rao, T. J. Sejnowski, Spike-timing-dependent Hebbian plasticity as temporal difference learning. *Neural Comput.* **13**, 2221–2237 (2001).
26. D. J. Foster, R. G. M. Morris, P. Dayan, A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus* **10**, 1–16 (2000).
27. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
28. Aidan Hogan *et al*., Knowledge graphs. *Syn. Lect. Data Semant. Knowl.* **12**, 1–257 (2021).
29. M. P. Lehmann *et al*., One-shot learning and behavioral eligibility traces in sequential decision making. *Elife* **8**, e47463 (2019).
30. A. Power, Y. Burda, H. Edwards, I. Babuschkin, V. Misra, Grokking: Generalization beyond overfitting on small algorithmic datasets. arXiv [Preprint] (2022). https://arxiv.org/abs/2201.02177 (Accessed 6 January 2022).
31. Z. Liu *et al*., Towards understanding grokking: An effective theory of representation learning. arXiv [Preprint] (2022). https://arxiv.org/abs/2205.10343 (Accessed 14 October 2022).
32. N. Frémaux, H. Sprekeler, W. Gerstner, Reinforcement learning using a continuous time actor-critic framework with spiking neurons. *PLoS Comput. Biol.* **9**, e1003024 (2013).
33. J. J. Hopfield, Neurodynamics of mental exploration. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 1648–1653 (2010).
34. T. Zhang, M. Rosenberg, P. Perona, M. Meister, Endotaxis: A universal algorithm for mapping, goal-learning, and navigation. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2021.09.24.461751 (Accessed 10 October 2022).
35. V. Mnih *et al*., Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).