



1 Background

A recent normative theory of hippocampal replay¹ suggests that the specific order of replay experiences is optimised for offline planning, whereby each replay corresponds to an update to a state-action value^{2,3}

The derived prioritisation scheme considers the expected improvement in the animal's immediately ensuing behaviour as a result of each potential individual replay update:

$$v_{\pi_{new}}(s) - v_{\pi_{old}}(s) = \underbrace{\sum_{i=0}^{\infty} \gamma^i P(s \rightarrow x, i, \pi_{old})}_{\text{Need}} \times \underbrace{\sum_a [\pi_{new}(a | x) - \pi_{old}(a | s)] q_{\pi_{new}}(x, a)}_{\text{Gain}} \quad \begin{array}{l} \pi_{old} \text{ Current policy [before replay update]} \\ \pi_{new} \text{ New policy [implied by a replay update]} \end{array}$$

2 Problem

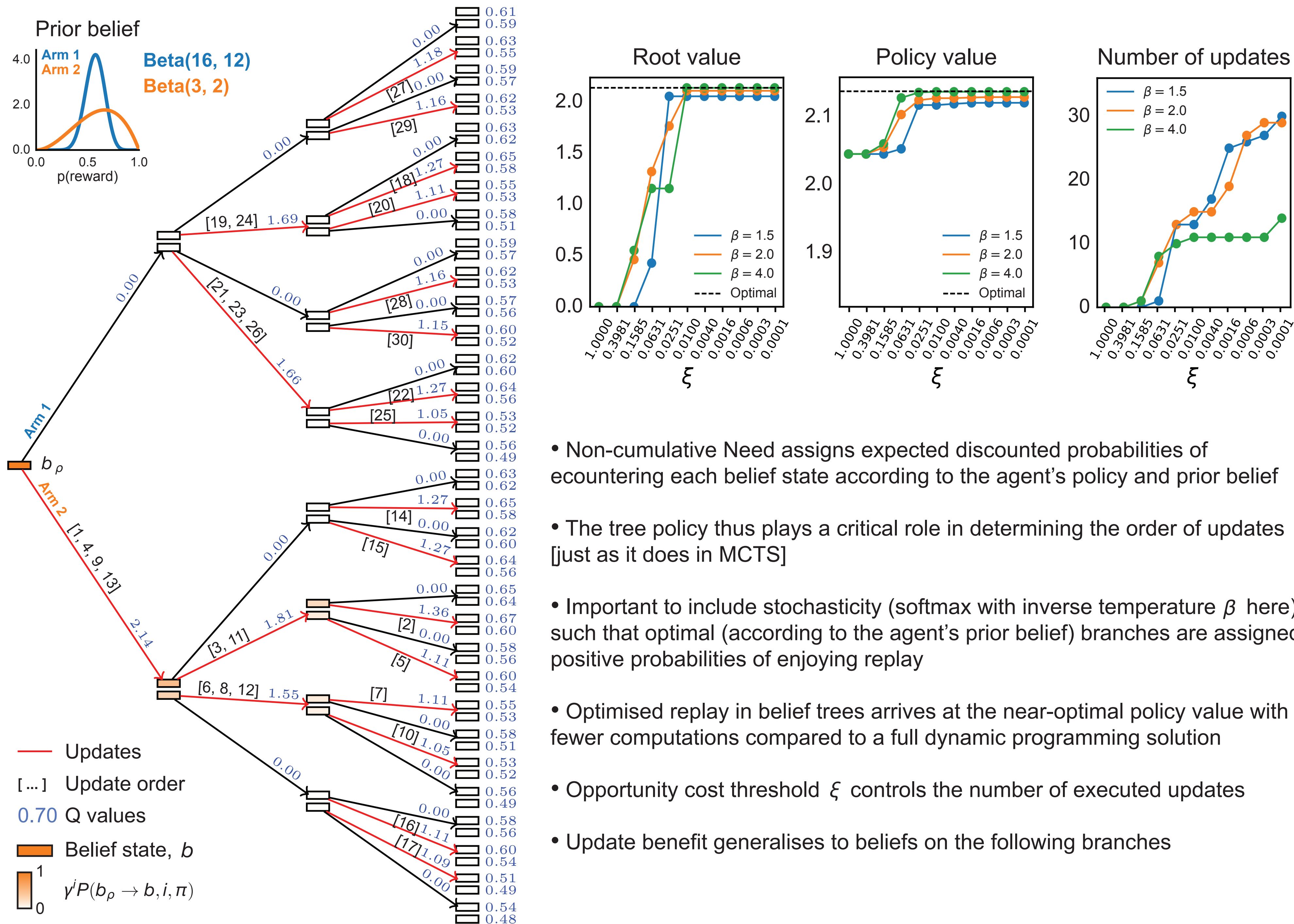
Although very successful, the theory does not extend to replay that encourages optimistic exploration in the face of ignorance

In this work, we extend the theory of replay to partially observable domains to show that optimised replay does favour uncertain outcomes whenever there is a potential long-range benefit of exploration

We consider optimised replay in belief MDPs where the following decomposition applies:

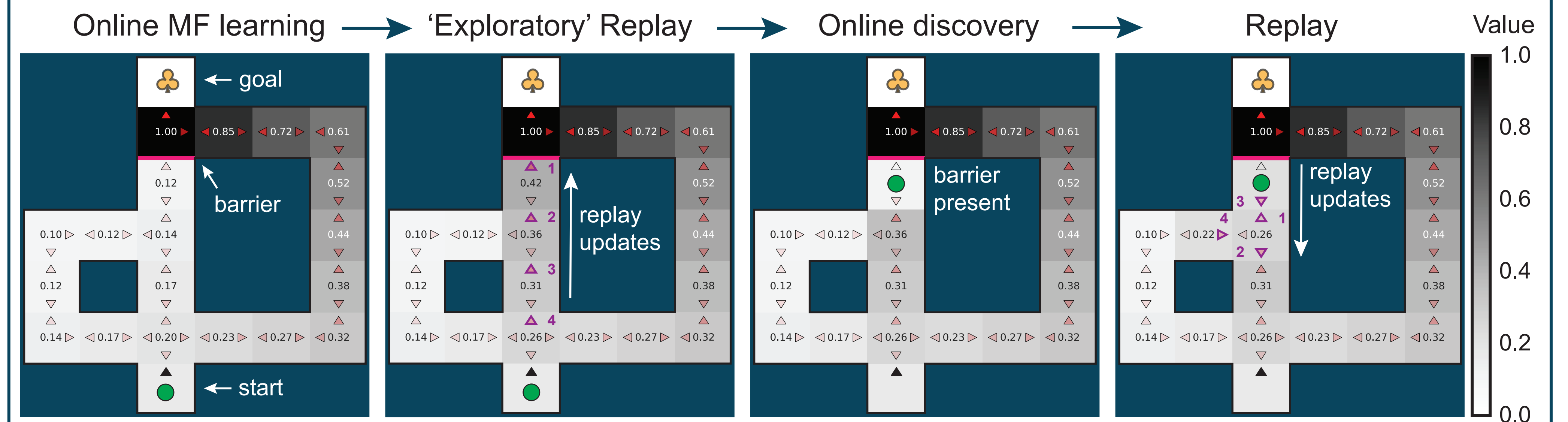
$$v_{\pi_{new}}(z) - v_{\pi_{old}}(z) = \sum_{z' \in \mathcal{Z}} \sum_{i=0}^{\infty} \gamma^i P(z \rightarrow z', i, \pi_{old}) \times \sum_a [\pi_{new}(a | z') - \pi_{old}(a | z')] q_{\pi_{new}}(z', a) \quad \text{where } z = \langle b \in \mathcal{B}, s \in \mathcal{S} \rangle$$

3 Prioritised sweeping in Bayesian bandits



- Non-cumulative Need assigns expected discounted probabilities of encountering each belief state according to the agent's policy and prior belief
- The tree policy thus plays a critical role in determining the order of updates [just as it does in MCTS]
- Important to include stochasticity (softmax with inverse temperature β here) such that optimal (according to the agent's prior belief) branches are assigned positive probabilities of enjoying replay
- Optimised replay in belief trees arrives at the near-optimal policy value with fewer computations compared to a full dynamic programming solution
- Opportunity cost threshold ξ controls the number of executed updates
- Update benefit generalises to beliefs on the following branches

4 Offline replay drives directed exploration

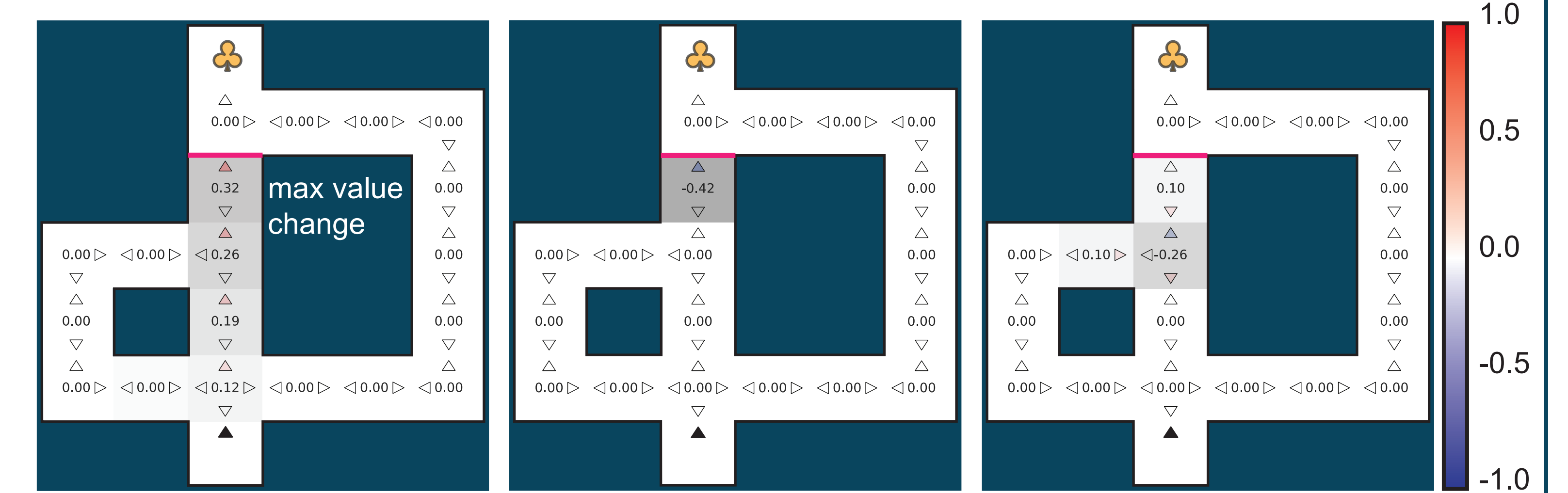


Value function changes occasioned by the above offline computations / behaviour

Tolman's⁴ detour maze treated as a POMDP

The agent is uncertain about the barrier configuration

Offline replay in belief space can give rise to directed exploration

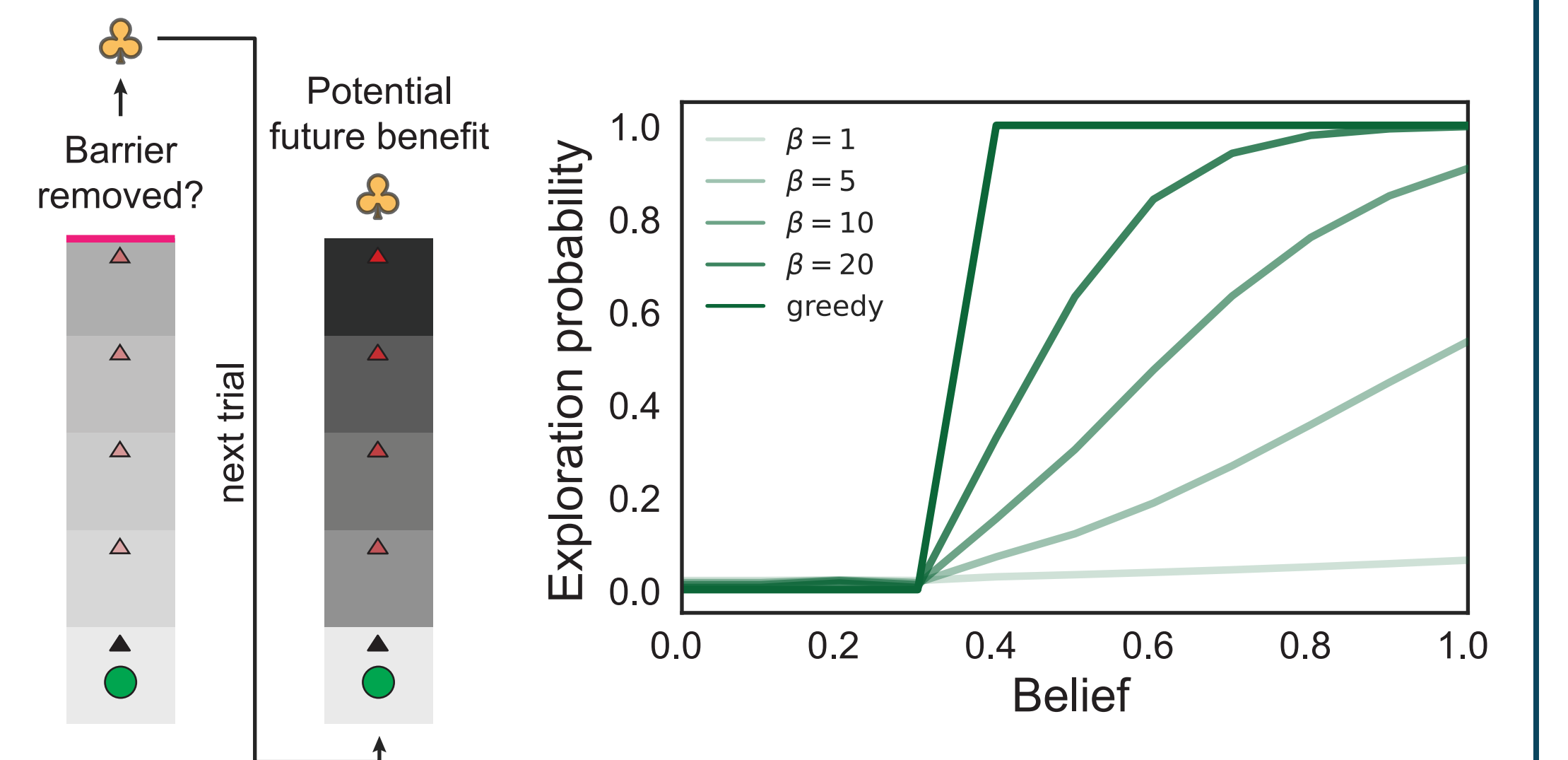


The agent maintains a belief over the presence of one of the barriers from the original Tolman's experiment

At the end of each trial, the belief is updated according to:

$$b_{t+1} = \begin{cases} \kappa \phi + (1 - \kappa) b_t & \text{if the shortcut was not attempted} \\ 1 - \kappa(1 - \phi) & \text{if the shortcut was successful} \\ \kappa \phi & \text{if the shortcut was unsuccessful} \end{cases}$$

and is set to what was actually experienced during the trial



5 Summary

- Belief-space formulation of optimised replay facilitates directed uncertainty-guided exploration
- Such replay propagates the value of the future use of potential information gain (deep exploration)
- Broad generalisation across belief states
- Predictions for the role of replay in rodent & human exploration

6 References

1. Mattar MG, Daw ND. Prioritized memory access explains planning and hippocampal replay. Nat Neurosci (2018).
2. Sutton RS. Dyna, an integrated architecture for learning, planning, and reacting. ACM Sigart Bulletin (1991).
3. Moore AW, Atkeson CG. Prioritized sweeping: Reinforcement learning with less data and less time. Machine learning (1993)
4. Tolman EC. Cognitive maps in rats and men. Psychological review (1948)

GA and PD are funded by the Max Planck society. PD is also funded by the Humboldt Foundation