We'd like to thank the reviewers for their encouraging comments and helpful suggestions on how to best improve the paper. We've tried to follows these as closely as possible and are excited to share the details below.

As suggested by both Reviewers 1&2 (**R1Q4&R2Q2**), we agree that it is much better to search through exploration and learning rates when feasible. We now show results for SR learning rate $\alpha_{SR} \in [0.001, 0.005, 0.01, 0.05, 0.1]$ and exploration $\epsilon \in [0., 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35]$ for Experiments 1&2. For Exp.1 BSR performed best in all 40 settings. Best-performing results (for total steps taken in 4500 episodes) are presented in Table I. We also agree with (**R1Q3&R2Q1**) that it is instructive to present per-episode statistics and trajectories, to make the results more interpretable and tie them in with existing literature, and thank **R1&2** for these suggestions. Per-episode steps are presented at the beginning, middle and end of training for respective best performing α_{SR} and ϵ (Fig. R1a), as well as an 'average' setting of $\alpha_{SR} = 0.01$ and $\epsilon = 0.1$ (Fig. R1b) that shows how SSR and GPI don't have the capacity to keep adapting in this case. For Exp.2 we also ran each parameter setting with or without CR offset based exploration (**R2Q3**) for BSR, SSR and EW, and added $\epsilon = 0.4$. Total (Table II) and per-episode returns (Fig. R1c) of the best parameter settings illustrate the effect of CR based exploration. Fig. 1Rd shows representative trajectories. We will

Table I

Algo	Steps(10^3)	ϵ, α_{SR}
BSR-4	348 ± 7	0., 0.005
GPI-4	400 ± 10	0.05, 0.001
SSR	398 ± 9	0.1, 0.001
KQ	385 ± 9	0.05, 0.001

Table II

	CR offset exploration		No CR offset exploration	
Algo	Returns(10^3)	ϵ, α_{SR}	Returns(10^3)	ϵ, α_{SR}
BSR-4	237 ± 4	0.25, 0.05	105 ± 4	0.3, 0.1
SSR	201 ± 4	0.25,0.005	87 ± 4	0.4, 0.001
EW-4	202 ± 4	025,0.05	78 ± 5	0.4, 0.005
Stored reward map				
GPI-4	196 ± 1	0.35, 0.01		

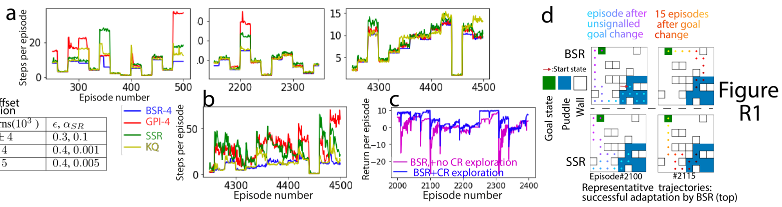

Figure R1

integrate these results into Fig. 2 and follow up with Exp. 3. We apologize for omitting important related work form our original submission (**R2Q1&Q4**), we now reference [1,2,3] when discussing limitations of SR for transfer in Sections 1, 4, and 6: In particular, we refer to the experiment in [1] showing the limited policy revaluation capabilities, and discuss how [2] finds evidence of these limitations in human behaviour. We also outline important, qualitative differences between the transfer experiments in [3] and our Experiment 1: in [3] the agent only has to learn four, partially disjoint trajectories (connecting opposite corners of an open maze), resulting in much more limited ambiguity in optimal action choice for most states. In signalled settings like our known quartile (KQ), or for agents with memory (e.g. RNNs) this becomes a simple task. In our case most states can be part of a large number of optimal trajectories, with internal walls providing for non-trivial dynamics. Studying performance across all these possible trajectories frames this problem properly in terms of lifelong/multitask learning, and highlights which algorithms can adapt well across all tasks.

We apologise to **R3** for any lack of clarity in our presentation and model. We aimed to use notation standard in the literature, but agree that it is crucial to clearly define all notation. We have reworked Section 3 to clarify the algorithmic components, added detailed figure legends (e.g. for 1b: Dirichlet process mixture model of the convolved reward maps. The model is defined by a base distribution H and concentration parameter α, giving a random distribution over CR maps.) and defined all notation. We now explicitly describe the Chinese restaurant process (CRP) view of the DP and our particle filter to clarify the inference process. Briefly,the CRP gives a closed-form prior over the discrete latent contexts at every step, each associated with both a CR map (e.g. CR_3 for context 3) and a successor map M. We base our inference on observed CR values that depend on the current reward function and the policy the agent is following according to the successor maps. This inference is intractable and we want to avoid specifying priors over M and CR maps (priors normally defined by H). Amortizing the inference, we calculate a posterior over latent contexts by combining the CRP prior with Gaussian likelihoods of the observed CR value given the value stored in the CR maps. We update the sampled successor map M independently by TD update (L101), and then update the CR map at the end of the episode. This allows the agent to both integrate evidence and refine the SR at each step, while updating the CR map of the overall most likely context. Regarding **R3Q1**, the subscript was indeed omitted by mistake from the SR learning rate α_{SR}, we corrected this and included a definition in the text. Similarly (**R3Q3**), the state embedding $\phi(s)$ and the reward vector **w** are now defined straight away, rather than later, and we point out that in the tabular case $\phi(s)$ is a one-hot encoding of states and **w** the corresponding vector of reward per state. On the suggestion of **R1Q1** we also include a detailed methodological description in the SI and the suggested reference. **R1Q2** raises an important question regarding biological plausibility. In the tabular case, our algorithm relies only on TD updates, delta rule/Rescorla-Wagner type update of the CR maps, and Bayesian filtering (filtering is implemented by neurons e.g in probabilistic population codes or sampling based methods[4]). In the continuous setting we do rely on backpropagating the TD error through an MLP, but see [5] for recent advances. We do believe this puts the algorithm firmly in the biologically plausible setting compared to approaches relying on recurrent policy gradients or meta-learning with long-range backprop through time. As suggested (**R1Q6**) there is also behavioural evidence that GPI is not a good fit in Section 5.2, as it both makes many error trials, and takes longer to reach the goal, while animals (and BSR) have few error trials and tend to head straight to a goal after initial mistakes. This is suggested by Fig.4h and discussed briefly, but we will include a more detailed discussion. The correlation coefficient in section 5.1 (**R1Q5**) measures the correlation between the episode number and the z-score difference (Fig. 3c). It gives a measure of the transitioning from the old to the new maps.

[1] Russek et al. 2017, [2]Momennejad et al. 2017, [3]Lehnert et al. 2017, [4]Kutschireiter et al 2017, [5] Whittington et al. [2019]