Emergent social transmission of model-based representations without inference

Silja Keßler[1] **(silja.kessler@student.uni-tuebingen.de), Miriam Bautista-Salinero**[1]**, Claudio Tennie**[1]
& Charley M. Wu[2,3] **(charley.wu@tu-darmstadt.de)**
[1] University of Tübingen, Tübingen, Germany
[2] Technical University Darmstadt, Darmstadt, Germany
[3] Hessian AI, Darmstadt, Germany

Abstract

How do people acquire rich, flexible knowledge about their environment from others despite limited cognitive capacity? Humans are often thought to rely on computationally costly mentalizing, such as inferring others' beliefs. In contrast, cultural evolution emphasizes that behavioral transmission can be supported by simple social cues. Using reinforcement learning simulations, we show how minimal social learning can indirectly transmit higher-level representations. We simulate a naïve agent searching for rewards in a reconfigurable environment, learning either alone or by observing an expert—crucially, without inferring mental states. Instead, the learner heuristically selects actions or boosts value representations based on observed actions. Our results demonstrate that these cues bias the learner's experience, causing its representation to converge toward the expert's. Model-based learners benefit most from social exposure, showing faster learning and more expert-like representations. These findings show how cultural transmission can arise from simple, non-mentalizing processes exploiting asocial learning mechanisms.

Keywords: Interactive behavior; Social cognition; Agent-based Modeling; Computational Modeling

Introduction

Reinforcement learning (RL) offers a powerful framework for understanding how agents acquire knowledge through interaction with their environment (Sutton & Barto, 2018). Traditionally, this framework has focused on asocial learning: individual agents learn by interacting with their environment and updating internal representations based on experience. While an asocial framing has yielded substantial advances (Shteingart & Loewenstein, 2014; O'Doherty, Lee, & McNamee, 2015), it stands in contrast to how humans and other animals leverage *social* information (Laland, 2004; Olsson, Knapska, & Lindström, 2020) to accelerate learning (Rendell, Boyd, et al., 2010; Park, Goïame, O'Connor, & Dreher, 2017), reduce exploration costs (Witt, Toyokawa, Lala, Gaissmaier, & Wu, 2024; Garg, Kello, & Smaldino, 2022), and improve performance in complex environments (Wu et al., 2025; Hawkins et al., 2023).

Thus, recent work in RL has increasingly incorporated social learning mechanisms, to both understand human social cognition (Wu et al., 2025; Hackel, Mende-Siedlecki, & Amodio, 2020; Najar, Bonnet, Bahrami, & Palminteri, 2020) and for engineering applications where learning from the environment is costly (Prasad et al., 2024; Urain et al., 2025). Learning from others is often modeled as *mentalizing* (Baker, Saxe, & Tenenbaum, 2009; Jara-Ettinger, 2019)—inferring latent mental states such as goals, preferences, or beliefs about the structure of the environment from observed actions (also known as Theory of Mind; Tomasello, 1999). Techniques like inverse reinforcement learning (IRL; Russell, 1998; Jara-Ettinger, 2019; Jern, Lucas, & Kemp, 2017; Collette, Pauli, Bossaerts, & O'Doherty, 2017) formalize this process by inverting the RL model to "unpack" observed behavior into underlying value or model-based representations of the world. Although powerful, such social inference methods are computationally costly and intractable in many settings (Wu, Vélez, & Cushman, 2022).

In contrast, cultural evolution research has long emphasized how sophisticated patterns of behavioral transmission can emerge from low-cost social cues (Heyes, 2018; Mesoudi, 2016). Individuals often rely on simple heuristics, such as selectively imitating successful (McElreath et al., 2008) or prestigious individuals (Jiménez & Mesoudi, 2019; Boyd & Richerson, 1988) without needing to infer mental states. These mechanisms can nonetheless support the accumulation of complex behaviors across generations (Legare & Nielsen, 2015; Rendell, Fogarty, & Laland, 2010), suggesting an alternative pathway for effective social learning without the costly inference required for mentalizing (Roberts-Gaal, Bolic, & Cushman, 2025; Uchiyama, Tennie, & Wu, 2023).

Here, we explore how higher-level representations of value and model-based state transitions can emerge in RL agents from simple forms of social learning—without explicit mentalizing—and how they support generalization to novel conditions. We simulate agents that search for rewards in a spatially reconfigurable environment, learning either independently or by observing an expert. Crucially, learners do not attempt to infer the expert's goals or internal model of the environment. Instead, they rely on simple social cues, adopting strategies more similar to imitation (Byrne & Russon, 1998) or local enhancement (Galef, 1988) than mentalizing. Specifically, we consider two strategies at different representational levels: *decision biasing* (DB; Toyokawa, Whalen, & Laland, 2019), which biases the agent's policy, and *value shaping* (VS; Najar et al., 2020), which shapes the agent's expected values for actions based on the expert's behavior. Both strategies are applied to model-free and model-based RL agents, either guiding learning either through model-free habit formation or via model-based planning (Drummond & Niv, 2020; Hackel, Berg, Lindström, & Amodio, 2019).

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

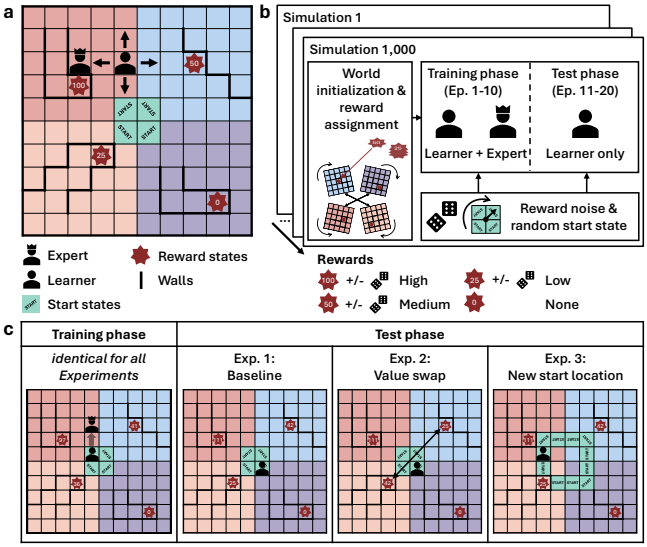

Figure 1: **Simulations**. **a**) Grid-world environment composed of four quadrants with fixed walls and designated reward states. **b**) For each simulation, quadrants were randomly rotated and arranged, while rewards were randomly assigned to designated reward states. The simulations comprised a training phase with an observable expert agent (overtrained model-based RL), followed by a test phase without the expert. A different start position was randomly selected on each episode (from the four central states), and stochastic noise was added to reward observations (see Methods). **c**) Experimental manipulations applied during the test phase: no change (Exp. 1), swap of two randomly selected rewards (Exp. 2), or shifted start positions (Exp. 3).

Across experiments, our results show that social learners—particularly those with model-based architectures—adopt more expert-like value and belief representations, supporting effective generalization to novel reward configurations and starting positions. These results demonstrate that simple, non-mentalizing social mechanisms can produce higher-level, flexible representations, highlighting a computational pathway through which complex cultural transmission can emerge from minimal social information.

Methods

We simulated agent behavior across three experiments ($n = 1{,}000$ per experiment) using a spatially reconfigurable environment, resembling a foraging-themed board game. Agents navigated generated grid-worlds over 20 episodes to acquire rewards. During training, agents learned through a combination of individual experience and observing an expert. Agent performance was then evaluated in the test phase, in which all agents continued to learn solely through individual experience, to assess generalization in the absence of the expert.

Environment and Rewards. The environment was composed of four quadrants, each forming a 5×5 grid of discrete states. Each quadrant contained a fixed configuration of walls blocking movement and a single designated reward state (Fig. 1a). In each simulation, we rotated and rearranged these quadrants to generate $4! \cdot 4^4 = 6{,}144$ unique configurations of 10×10 grid worlds. Reward values for each designated reward state were randomly assigned (without replacement) at the start of each simulation from four values: 0 (no reward), 25 (low), 50 (medium), and 75 (high). To introduce stochasticity, noise ε was applied independently to each reward observation using a shifted binomial distribution $\varepsilon \sim \text{Binomial}(n = 400, p = 0.5) - 200$, yielding integer values with a mean of 0 and a variance of 100.

Procedure. At the start of each episode, the agent was placed in one of four start states at the center of the board. The agent moved between adjacent states by executing actions (up, down, left, right), freely moving between quadrants, constrained by walls. Each attempts to move into a wall resulted in the agent remaining in the prior state. Actions not yielding a positive reward incurred a cost (negative reward) of -1. An episode terminated either when a positive reward was obtained or after 40 actions.

Simulations were divided into a *training phase* (episodes 1-10) and a *test phase* (episodes 11-20; Fig. 1b). During the training phase, learners were accompanied by an expert, which was an asocial model-based RL agent (see Fig. 2) that was pre-trained for 120 episodes in the same environment. Thus, the training phase offered social learning agents (defined below) the ability to combine information from the expert with knowledge gained through their own experiences. In contrast, asocial learners could only acquire knowledge through direct interactions with the environment. In the subsequent test phase, the expert was removed, meaning all learners could only learn asocially.

Experiments. We conducted three experiments to assess the learners' ability to generalize knowledge acquired during training (Fig. 1c). Exp. 1 established baseline performance and examined the transfer of values and model-based beliefs from the expert to social learners, without any changes to the environment. In Exps. 2 and 3, the test phase introduced additional modifications to the task to assess the learner's robustness to changes in the environment. Exp. 2 modified the reward structure, by swapping the values of two randomly selected reward states. Exp. 3 modified the starting locations, by shifting start states one tile towards the corners (excluding reward states), while keeping the spatial layout fixed.

Computational models

We compared six computational learning models (Fig. 2) with a 2×3 factorial design contrasting model-free vs model-based RL across three forms of social learning: asocial learning (AS), decision biasing (DB), and value shaping (VS). Asocial learners exclusively relied on individual experiences, whereas social learners additionally incorporated expert information either at the action selection level by biasing their policy toward the expert's behavior (DB), or at the value learning level by directly shaping value representations (VS) based on observed expert actions.

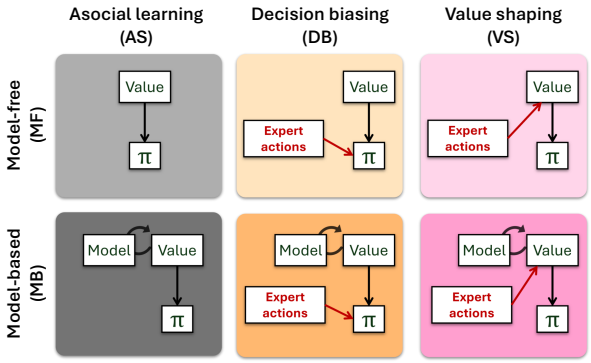

Figure 2: **Learning Strategies.** Six learning conditions were tested, combining model-free (MF) and model-based (MB) RL with either asocial learning (AS) or social learning from an expert. Social learning was implemented at two levels: policy-based (DB) and value-based (VS).

Model-free (MF) learning was implemented as Q-learning with Temporal Difference (TD) prediction updates:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a) \right) \tag{1}$$

where $Q(s,a)$ is the state-action value indicating the expected cumulative rewards for performing action a in state s. Q-values were initialized uniformly to 1 for all state–action pairs and updated based on the learning rate α and proportional to the difference between the prior Q-value and the sum of the reward obtained in the current state $R(s)$ and the maximum next-state Q-value $\max_{a'} Q(s',a')$ discounted by a temporal discount factor $\gamma \in [0,1]$.

Using these learned Q-values, the asocial learner computed a policy π_{asoc} based on a softmax function whose randomness was controlled by the inverse-temperature parameter β:

$$\pi_{asoc}(a|s) \propto \exp\left(\beta \cdot Q(s,a)\right) \tag{2}$$

Model-based (MB) learning was implemented using Dyna-Q (Sutton, 1990), a variant of Q-learning in which the agent maintains an internal model (i.e., beliefs) of the environment, defined as the transition probabilities between states given an action, $B(s'|s,a)$. Beliefs were updated via the Delta-rule (Antonov & Dayan, 2025):

$$B(s'|s,a) \leftarrow B(s'|s,a) + \eta \left(\delta(s',s) - B(s'|s,a) \right) \tag{3}$$

where the Kronecker delta $\delta(s',s) = 1$ if the transition $s \rightarrow s'$ is successful, and 0 otherwise.

These beliefs could be used to update Q-values more efficiently through simulated experiences (dependent on B), rather than solely relying on direct RL by executing actions in the world. This supports planning-based decision-making and is a cornerstone of modern theories of human learning (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Drummond & Niv, 2020; Wu et al., 2022). The following model-based planning loop was executed k times after each action, with $k \sim \text{Pois}(\lambda)$ sampled each time from a Poisson distribution to introduce stochasticity and λ is treated as a free parameter (planning rate). Q-values and beliefs were maintained across episodes, such that knowledge acquired during training carried over into the test phase.

Algorithm 1 Dyna-Q inner loop

for $i = 1, 2, \ldots, k$ **do**
  $S \leftarrow$ random previously observed state
  $A \leftarrow$ random action previously taken in S
  $R, S' \leftarrow B(S,A)$
  $Q(s,a) \leftarrow Q(s,a) + \alpha\left(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$
end for

Decision-biasing (DB) was inspired by algorithmic accounts of social learning in bandit tasks (Toyokawa et al., 2019; Witt et al., 2024), but modified for our sequential navigation task. We implemented DB by combining the asocial policy (Eq. 2) with a purely social policy π_{soc} during training, weighted by mixture parameter $\omega \in [0,1]$:

$$\pi_{\text{DB}} = (1-\omega)\pi_{asoc}(s,a) + \omega \pi_{soc}(s,a) \tag{4}$$

The social policy π_{soc} deterministically selects the action that minimizes the distance to expert:

$$\pi_{\text{soc}}(a \mid s) = \mathbb{I}\left[a = \arg\min_{a'} D\left(s^{t}_{\text{expert}}, s'(s,a')\right) \right] \tag{5}$$

where the indicator function $\mathbb{I}[\cdot]$ assigns 1 to the action that minimizes D, and 0 otherwise. Distance D corresponds to the expected number of actions needed to reach the expert's last observed state s^{t}_{expert} from the learner's current state s'. For MF agents, D is approximated using Manhattan distance (ignoring boundaries or walls, since MF agents do not represent environmental structure). For MB agents, D is computed via value iteration under the agent's beliefs: $D(s) = \min_a \sum_{s'} B(s'|s,a')(1 + D(s'))$. When the expert terminated prior to the learner, the expert's last visited location was frozen.

Value shaping (VS) was inspired by theoretical accounts of the local enhancement of stimuli based on observing others engaging with them (Spence, 1937; Galef, 1988). Following Najar et al. (2020), we implemented VS by augmenting the learner's value representations during training for actions performed by the expert in a given state:

$$Q(s_{expert}, a_{expert}) \leftarrow Q(s_{expert}, a_{expert}) + \kappa \mathbb{I}(s_{expert}, a_{expert}). \tag{6}$$

κ is a bonus parameter that directly shapes the learner's value function by enhancing state-action pairs observed in the expert. The boosted Q-values were then passed through a softmax (cf. Eq. 2), increasing the likelihood of selecting actions observed in the expert in the corresponding state. Once the expert reached a termination step, the VS learner reduced to asocial learning for the remainder of the episode.

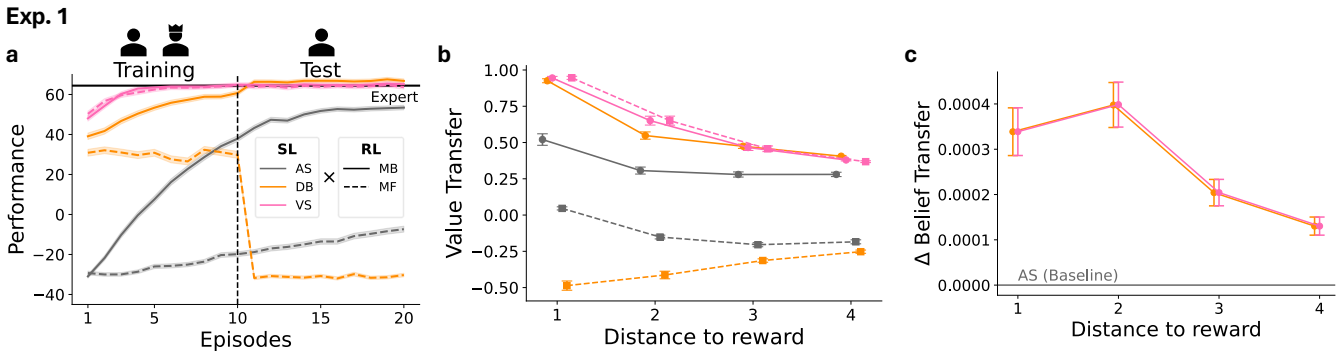

Figure 3: **Experiment 1.** **a**) Performance: mean cumulative reward for different learning strategies combining (a-)social learning (SL) and reinforcement learning (RL) strategies. The vertical dotted line represents the transition between training and test phase. The black horizontal line shows the expert's mean performance over episodes observed by the social learners. **b**) Value transfer: mean correlation between learner's and expert's value function, grouped by distance to reward states. **c**) Belief transfer: mean correlation between learner and expert transition matrices, grouped by distance to reward states. Correlations are shown for model-based social learners normalized relative to the asocial model-based learner (horizontal grey line). Shaded areas and error bars denote the standard error of the mean (SEM) across simulations.

Parameter optimization. We optimized each model's hyperparameters using the differential evolution algorithm (Storn & Price, 1997) to maximize performance (cumulative reward) across 1,000 independent simulations in the baseline environment. Optimization for social learners was based solely on the 10 training episodes, whereas asocial agents and the expert were optimized using all available episodes (20 and 120, respectively). Learning rates α and η for social learners were fixed to the values optimized for asocial learners, due to joint optimization consistently converging to low learning rates that suppressed individual learning and led to poor test performance. Candidate parameters were optimized in a transformed, unbounded space to enforce valid ranges (e.g., log-transform for $\beta \in [0, \infty]$) and a fixed random seed was used to ensure reproducibility.

Results

We first evaluated the performance of all agents across episodes, distinguishing between training and test phases to characterize how effectively agents acquired and applied knowledge, depending on their learning strategy. Next, we examined how information was indirectly transferred from expert to social learners by analyzing value and belief representations, assessing whether social learners acquired similar value and state transition representations as encoded by the expert. Finally, we tested the robustness of the learned knowledge under environmental changes: in Experiment 2 we assessed how well learners generalized to modified reward configurations, and in Experiment 3 we modified the starting position to evaluate whether the belief transferred from the expert was preserved and continued to support behavior under novel start conditions. Across analyses, variability across simulations was quantified using the standard error of the mean (SEM).

Exp. 1: Value/belief transfer without mentalizing

Exp. 1 established a baseline in which training and test phases were identical except for the expert's removal. Performance was quantified as the mean cumulative reward per learner across simulations, demonstrating how effectively different learning strategies enabled agents to navigate the environment and collect rewards over the course of the 20 episodes.

Starting with the asocial agents (AS), we observed a strong advantage of model-based (MB) over model-free (MF) learning (Fig. 3a). However, all social learning agents (DB and VS) outperformed the best asocial agent (AS-MB) in the training phase by leveraging information from the expert. Among social strategies, VS agents were superior, reaching expert-level performance with no difference between MB and MF. This reflects differences in how social information is incorporated. VS agents directly learn state-action values from the observed expert behavior, enabling effective generalization across the state space, whereas DB agents acquire values indirectly from expert trajectories, which confines learning signals to a narrow region and reduces flexibility. Moreover, combining social and individual information at the policy level can create conflict between following the expert and exploiting the DB agent's own experience. After the removal of the expert in the test phase, VS performance remained stable, demonstrating the robustness of value-based social learning—notably, also for MF. In contrast, DB agents diverged sharply: DB-MF performance collapsed below AS-MF, suggesting an over-reliance on the expert that produced an unreliable value function beyond the expert's trajectories. DB-MB, however, continued to improve and ultimately surpassed both VS agents, replanning more effectively once the bias was removed.

We then computed indirect *value transfer* based on the correspondence between the learner's and expert's state-action value representations at the end of each simulation (averaged across simulations). Specifically, we computed Spearman

correlations between value functions across states, grouped by distance from the nearest reward to account for goal-relevance (i.e., states closer to rewards have greater relevance). This assessed how effectively the learner reconstructs the expert's value information. Figure 3b shows that except for DB-MF, all social agents achieved greater value transfer compared to asocial baselines (AS-MB and AS-MF). While asocial agents were bound to acquire similar value representations as the expert due to interacting with the same environment, social agents (except for DB-MF) achieved a substantial boost, providing a mechanistic explanation for their better performance. This value transfer effect also displayed goal-relevance, with stronger transfer for states closer to rewards.

Lastly, we quantified indirect *belief transfer* for MB learners as the correspondence between learner and expert transition matrices at the end of each simulation. We computed Pearson correlations across states, grouped by distance from rewards, and normalized them relative to AS-MB as an asocial baseline (Fig. 3c). We observed positive belief transfer for both MB social learners (equivalent for DB-MB and VS-MB), indicating that they acquired an internal model of state transitions that more closely matches the expert's compared to the asocial baseline. This effect was also influenced by goal-relevance, with states closer to rewards corresponding to greater belief transfer to our model-based social learners.

Exp. 2: Robustness to changes in reward

In Exp. 2, the test phase additionally included swapping the values of two randomly selected reward states (Fig. 1c). As in Exp. 1, we quantified performance as the mean cumulative reward per learner across simulations to assess how effectively different learning strategies generalized to the altered reward contingencies, focusing on the test phase since the training phase was identical to Exp. 1.

Test performance initially dropped for all agents (except AS-MF), reflecting the challenge of adapting to new reward contingencies—specifically for VS agents whose boosted values were now potentially misleading. The most pronounced drop occurred for DB-MF, which showed a similar collapse as in Exp. 1. However, all MB agents adapted to the new reward configuration and recovered performance.

To examine the underlying mechanisms, we analyzed *value accuracy* by computing the Spearman correlation between each learner's final value function (averaged across simulations) and the optimal value function for the new reward configuration, derived from the Bellman (1957) equation:

$$Q^*(s,a) = \sum_{s'} P(s' \mid s,a)\left[R(s,a,s') + \gamma \max_{a'} Q^*(s',a')\right] \quad (7)$$

with $\gamma = 0.99$. Here, we only included states visited at least once by the learner to avoid distortions from unvisited states. Correlations were again grouped by distance from the nearest reward to account for goal-relevance.

Our results show that social learning facilitates better generalization to changes in reward structure, with all social agents achieving greater value accuracy compared to asocial

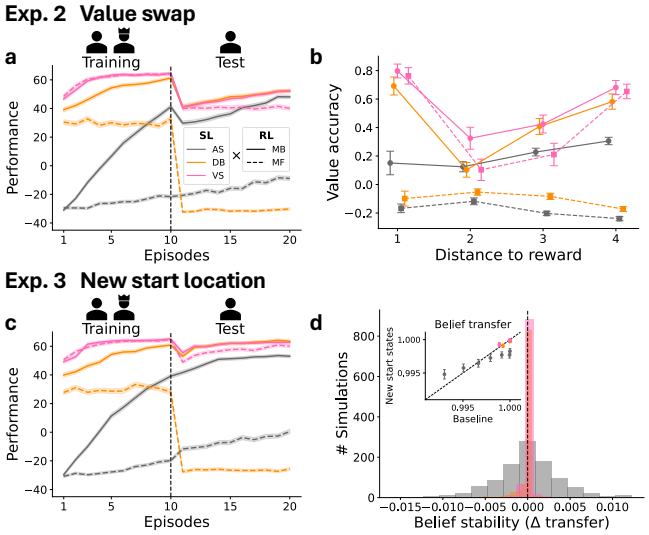

Figure 4: **Exps. 2 and 3.** **a**) Exp. 2 performance. **b**) Exp. 2 value accuracy: mean correlation between the learner's and the optimal value function for the modified reward structure (Eq. 7), grouped by distance to reward. **c**) Exp. 3 performance. **d**) Exp. 3 belief stability (MB only). Deviation in belief transfer between baseline and new start locations in the test phase across simulations; zero (dashed line) indicates robust preservation of model-based representations. Inset: comparison of belief transfer baseline (x-axis) vs. new start location (y-axis); points represent binned averages across simulations (quantiles) and error bars indicate 95% confidence intervals. The dashed diagonal line shows $y = x$.

baselines (AS-MB and AS-MF), except for DB-MF, whose poorly aligned value representations explain its poor performance (Fig. 4b). Notably, accuracy dropped at $d = 2$ for all agents except AS-MF and DB-MF, likely because one reward state was reachable via alternative paths diverging at that point (Fig. 1a). Further analyses indicated that agents consistently committed to one of the routes and as a result did not update the values of the unchosen, equally optimal path. AS-MF and DB-MF, in contrast, show a slight increase in accuracy at $d = 2$. These agents were more exploratory and therefore acquired broader knowledge about regions surrounding rewards. However, they more frequently settled for suboptimal rewards, resulting in lower values for states directly preceding a reward ($d = 1$).

Exp. 3: Robustness to changes in starting location

In Exp. 3, we tested robustness to changes in the starting positions (Fig. 1c). Focusing on the test phase, asocial learners continued to perform well, showing little sensitivity to the new start positions. All social learners (except for DB-MB) experienced a small initial dip in performance, but rapidly recovered to training-level performance. DB-MF suffered again a severe collapse similar to Exps. 1 and 2. These results reinforce the finding that policy-based social learning is fragile under environmental changes unless paired with a model-based system. Importantly, model-based social learners (DB-

MB and VS-MB) maintained a clear performance advantage over the asocial baseline (AS-MB) throughout the test phase, highlighting the robustness of model-based social strategies.

To examine whether beliefs acquired by model-based agents via the expert's actions were preserved under novel starting positions (Fig. 4d), we quantified belief stability as the deviation in indirect belief transfer (correlation between agent and expert transition matrices, grouped across distances) between Exp. 1 (baseline start positions) versus Exp. 3 (new start positions). Social learners, particularly VS-MB, showed notably less deviation than AS-MB, indicating that their reconstructed beliefs were largely preserved and remained highly similar to the expert (Fig. 4d inset). This showcases that social learners acquired more accurate, robust and presumably more global beliefs by leveraging the expert's knowledge—resulting in internal representations that generalize more effectively across starting positions.

Overall, these results demonstrate that particularly model-based social learners not only acquire similar structural knowledge to the expert, but also exhibit improved performance that remains robust under environmental changes—all without mentalizing.

Discussion

We investigated how rich, flexible knowledge can be socially transmitted without explicit mentalizing, thus sidestepping costly inferences about others' goals or beliefs. Using simulations with both model-free (MF) and model-based (MB) reinforcement learning (RL) agents, we show that learners can acquire higher-level representations through simple social heuristics. Specifically, decision-biasing (DB) operates at the policy-level, minimizing the distance to an expert demonstrator (Eqs. 5-4), while value shaping (VS) adds a value bonus to state-action pairs observed in the expert, akin to local enhancement (Eq. 6). Across three sets of simulations in a spatial foraging task, we found that social learning particularly benefited MB agents, which acquired more accurate and robust representations, while also generalizing better to changes in rewards and starting locations.

What mechanisms underlie these results? Even simple social heuristics bias the learner's experience, which, when leveraged by model-based planning, can support more accurate representations of value and environment structure (Wu et al., 2022; Uchiyama et al., 2023). In our simulations, MB agents use Dyna (Alg. 1) to simulate experiences and propagate values, analogous to hippocampal replay (Ólafsdóttir, Bush, & Barry, 2018; Miller, Botvinick, & Brody, 2017; Vikbladh et al., 2019). This suggests that insight can emerge from naïve imitation (Lyons, Young, & Keil, 2007) by exploiting planning mechanisms. This account is consistent with neural evidence that social observations evoke replay (Mou & Ji, 2025; Clein & Gould, 2025), pointing to a shared neural basis for social and asocial model-based learning.

Across both MB and MF agents, value-based social learning (VS) generally outperformed policy-based methods (DB), echoing a general principle in RL that value is a more generalizable representation than actions (Lehnert, Littman, & Frank, 2020; Wu et al., 2022). Notably, VS maintained strong performance even under model-free learning (VS-MF). In contrast, policy-based social learning was fragile under model-free learning (DB-MF), but robust when paired with model-based mechanisms (DB-MB). A likely explanation is that DB agents overfit to the expert's demonstrated trajectories, resulting in limited exploration of the state space. When the expert is removed, DB-MF agents are left with an underdeveloped value function and no mechanism to recover. DB-MB agents, however, can use their learned model to replan and update their value representation effectively.

These findings offer a potential perspective on cross-species differences between humans and non-human primates (NHP; Bandini & Tennie, 2023; Call, Carpenter, & Tomasello, 2005). Humans over-imitate, copying even irrelevant actions (Horner & Whiten, 2005; Lyons et al., 2007), whereas chimpanzees tend to emulate goals rather than copy specific behavior (Tennie, Call, & Tomasello, 2010). While the literature on model-based learning in NHPs is scarce (but see Sato, Sakai, & Hirata, 2023, for a task disincentivizing MB learning), differences in reliance on model-based planning may shift the cost–benefit trade-off between policy-based and value-based social learning (Wu et al., 2022), which could explain why over-imitation is more adaptive in humans than in chimpanzees.

Despite promising results, there are several limitations that could be addressed in future work. First, while we evaluated robustness to changes in rewards and starting locations, we did not manipulate the underlying environmental structure itself (e.g., changes to transition dynamics or wall layouts), which would allow us to better assess whether non-mentalizing social learning also supports the transfer of deeper structural knowledge (Ho, Saxe, & Cushman, 2022). Second, we did not incorporate mentalizing agents, e.g., agents using inverse reinforcement learning (IRL; Russell, 1998) to infer the expert's values or beliefs, into our simulation. Including such agents would provide an important comparative baseline, enabling more direct assessment of trade-offs between computationally costly mentalizing and simpler social heuristics like DB and VS (Devaine, Hollard, & Daunizeau, 2014). Third, we modeled the expert as a passive demonstrator rather than providing pedagogical demonstrations (Vélez, Chen, Burke, Cushman, & Gershman, 2023). Future work could test how such demonstrations might interact with or amplify our minimal social learning mechanisms. Finally, our results are based on simulations, whereas testing these predictions in human behavioral experiments will be essential to assess whether similar non-mentalizing processes can also support higher-order social transmission in people.

In sum, we provide a computational account of how the indirect transmission of high-level representations can arise from minimal, non-mentalizing forms of social learning that exploit asocial model-based RL mechanisms.

Acknowledgments

We thank Ryutaru Uchiyama and members of the HMC Lab for helpful discussions in the development of this project. This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (C^4: 101164709), the Hessian research funding programme LOEWE/4b//519/05/01.002(0022)/119, the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223), and the Excellence Cluster "Reasonable AI" by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy – EXC-3057.

References

Antonov, G., & Dayan, P. (2025). Exploring replay. *Nature Communications*, *16*(1), 1657.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. doi: 10.1016/j.cognition.2009.07.005

Bandini, E., & Tennie, C. (2023). Naïve, adult, captive chimpanzees do not socially learn how to make and use sharp stone tools. *Scientific Reports*, *13*(1), 22733.

Bellman, R. (1957). *Dynamic programming* (1st ed.). Princeton, NJ, USA: Princeton University Press.

Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago press.

Byrne, R. W., & Russon, A. E. (1998, October). Learning by imitation: A hierarchical approach. *Behavioral and Brain Sciences*, *21*(5), 667–684. doi: 10.1017/S0140525X98001745

Call, J., Carpenter, M., & Tomasello, M. (2005). Copying results and copying actions in the process of social learning: chimpanzees (pan troglodytes) and human children (homo sapiens). *Animal cognition*, *8*(3), 151–163.

Clein, R. S., & Gould, E. (2025). Representations of social experience in hippocampal circuits. *Trends in Cognitive Sciences*.

Collette, S., Pauli, W. M., Bossaerts, P., & O'Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *Elife*, *6*, e29718.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Devaine, M., Hollard, G., & Daunizeau, J. (2014). Theory of mind: did evolution fool us? *PloS One*, *9*(2), e87619.

Drummond, N., & Niv, Y. (2020). Model-based decision making and model-free learning. *Current Biology*, *30*(15), R860–R865.

Galef, B. G., Jr. (1988). Imitation in animals: History, definition and interpretation of the data from the psychological laboratory. In T. R. Zentall & B. G. Galef Jr (Eds.), *Social learning: Psychological and biological perspectives* (pp. 15–40). Psychology Press.

Garg, K., Kello, C. T., & Smaldino, P. E. (2022). Individual exploration and selective social learning: balancing exploration–exploitation trade-offs in collective foraging. *Journal of the Royal Society Interface*, *19*(189), 20210915.

Hackel, L. M., Berg, J. J., Lindström, B. R., & Amodio, D. M. (2019). Model-based and model-free social cognition: investigating the role of habit in social attitude formation and choice. *Frontiers in psychology*, *10*, 2592.

Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology*, *88*, 103948.

Hawkins, R. D., Berdahl, A. M., Pentland, A. S., Tenenbaum, J. B., Goodman, N. D., & Krafft, P. (2023). Flexible social inference facilitates targeted social learning when rewards are not observable. *Nature Human Behaviour*, *7*(10), 1767–1776.

Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*, *26*(11), 959–971.

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (pan troglodytes) and children (homo sapiens). *Animal cognition*, *8*(3), 164–181.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Jiménez, Á. V., & Mesoudi, A. (2019). Prestige-biased social learning: Current evidence and outstanding questions. *Palgrave Communications*, *5*(1).

Laland, K. N. (2004, February). Social learning strategies. *Learning & Behavior*, *32*(1), 4–14. doi: 10.3758/bf03196002

Legare, C. H., & Nielsen, M. (2015). Imitation and innovation: The dual engines of cultural learning. *Trends in cognitive sciences*, *19*(11), 688–699.

Lehnert, L., Littman, M. L., & Frank, M. J. (2020). Reward-predictive representations generalize across tasks in reinforcement learning. *PLoS computational biology*, *16*(10), e1008317.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, *104*(50), 19751–19756.

McElreath, R., Bell, A. V., Efferson, C., Lubell, M., Richerson, P. J., & Waring, T. (2008). Beyond existence and aiming outside the laboratory: estimating frequency-dependent and pay-off-biased social learning strategies. *Philosophical Transactions of the Royal Society B: Biological Sciences*,

363(1509), 3515–3528.

Mesoudi, A. (2016). Cultural evolution: a review of theory, findings and controversies. *Evolutionary biology*, *43*(4), 481–497.

Miller, K. J., Botvinick, M. M., & Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nature neuroscience*, *20*(9), 1269–1276.

Mou, X., & Ji, D. (2025). Observational activation of anterior cingulate cortical neurons coordinates hippocampal replay in social learning. *Elife*, *13*, RP97884.

Najar, A., Bonnet, E., Bahrami, B., & Palminteri, S. (2020). The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLoS biology*, *18*(12), e3001028.

Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The role of hippocampal replay in memory and planning. *Current Biology*, *28*(1), R37–R50.

Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Reviews Neuroscience*, *21*(4), 197–212.

O'Doherty, J. P., Lee, S. W., & McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, *1*, 94–100.

Park, S. A., Goïame, S., O'Connor, D. A., & Dreher, J.-C. (2017). Integration of individual and social information for decision-making in groups of different sizes. *PLoS Biology*, *15*(6), e2001958.

Prasad, V., Kshirsagar, A., Koert, D., Stock-Homburg, R., Peters, J., & Chalvatzaki, G. (2024). Moveint: Mixture of variational experts for learning human–robot interactions from demonstrations. *IEEE Robotics and Automation Letters*, *9*(7), 6043–6050.

Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., … Laland, K. N. (2010). Why copy others? insights from the social learning strategies tournament. *Science*, *328*(5975), 208–213.

Rendell, L., Fogarty, L., & Laland, K. N. (2010). Rogers'paradox recast and resolved: Population structure and the evolution of social learning strategies. *Evolution*, *64*(2), 534–548.

Roberts-Gaal, X., Bolic, M., & Cushman, F. A. (2025). Environmental variability shapes the representational format of cultural learning. *Proceedings of the National Academy of Sciences*, *122*(28), e2505283122.

Russell, S. (1998, July). Learning agents for uncertain environments. In (p. 101–103). Madison Wisconsin USA: ACM. doi: 10.1145/279943.279964

Sato, Y., Sakai, Y., & Hirata, S. (2023). State-transition-free reinforcement learning in chimpanzees (pan troglodytes). *Learning & behavior*, *51*(4), 413–427.

Shteingart, H., & Loewenstein, Y. (2014). Reinforcement learning and human behavior. *Current opinion in neurobiology*, *25*, 93–98.

Spence, K. W. (1937, December). Experimental studies of learning and the higher mental processes in infra-human primates. *Psychological bulletin*, *34*(10), 806–850. doi: 10.1037/h0061498

Storn, R., & Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*(4), 341–359. doi: 10.1023/A:1008202821328

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the seventh international conference on machine learning* (p. 216–224).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tennie, C., Call, J., & Tomasello, M. (2010). Evidence for emulation in chimpanzees in social settings using the floating peanut task. *PLoS One*, *5*(5), e10544.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA, US: Harvard University Press. (Pages: vi, 248)

Toyokawa, W., Whalen, A., & Laland, K. N. (2019, January). Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, *3*(2), 183–193. doi: 10.1038/s41562-018-0518-x

Uchiyama, R., Tennie, C., & Wu, C. M. (2023). Model-based assimilation transmits and recombines world models. In L. Hunt, C. Summerfield, T. Konkle, E. Fedorenko, & T. Naselaris (Eds.), *Proceedings of the 2023 Conference on Cognitive Computational Neuroscience.* Oxford, UK. doi: 10.31234/osf.io/v69jy

Urain, J., Mandlekar, A., Du, Y., Muhammad, N., Xu, D., Fragkiadaki, K., … others (2025). A survey on deep generative models for robot learning from multimodal demonstrations. *IEEE Transactions on Robotics*, *42*, 60–79.

Vélez, N., Chen, A. M., Burke, T., Cushman, F. A., & Gershman, S. J. (2023). Teachers recruit mentalizing regions to represent learners' beliefs. *Proceedings of the National Academy of Sciences*, *120*(22), e2215015120.

Vikbladh, O. M., Meager, M. R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., … Daw, N. D. (2019). Hippocampal contributions to model-based planning and spatial memory. *Neuron*, *102*(3), 683–693.

Witt, A., Toyokawa, W., Lala, K. N., Gaissmaier, W., & Wu, C. M. (2024). Humans flexibly integrate social information despite interindividual differences in reward. *Proceedings of the National Academy of Sciences*, *121*(39), e2404928121. doi: 10.1073/pnas.2404928121

Wu, C. M., Deffner, D., Kahl, B., Meder, B., Ho, M. K., & Kurvers, R. H. (2025). Adaptive mechanisms of social and asocial learning in immersive foraging environments. *Nature Communications*, *16*, 3539. doi: 10.1038/s41467-025-58365-6

Wu, C. M., Vélez, N., & Cushman, F. A. (2022). Representational exchange in human social learning: Balancing efficiency and flexibility. In I. C. Dezza, E. Schulz, & C. M. Wu (Eds.), *The Drive for Knowledge: The Science*

of Human Information-Seeking. Cambridge: Cambridge University Press.