
Article:
Barfuss, W., Tittel, P. and Mann, R.P. (Accepted: 2026) Learning Together, Better, Faster - On the Timescales of Learning Social Learning Strategies. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. ISSN: 1364-503X (In Press)

LEARNING TOGETHER, BETTER, FASTER

ON THE TIMESCALES OF LEARNING SOCIAL LEARNING STRATEGIES

Wolfram Barfuss
Transdisciplinary Research Area Sustainable Futures, University of Bonn, 53115 Bonn, Germany
Center for Development Research, University of Bonn, 53113 Bonn, Germany
Institute for Food & Resource Economics, University of Bonn, 53115 Bonn, Germany
wbarfuss@uni-bonn.de

Paula Tittel
Faculty of Mathematics and Physics, Leibniz University Hannover, 30167 Hannover, Germany

Richard P. Mann
Department of Statistics, School of Mathematics, University of Leeds, Leeds, LS2 9JT, United Kingdom

April 24, 2026

ABSTRACT

Social learning strategies are the foundational mechanism for human-machine cultural evolution. For decades, they have been studied either as fixed evolutionary heuristics or as rational Bayesian solutions, yet neither approach explains their remarkable flexibility nor how boundedly rational agents could acquire them. Recent work has begun to explain social learning as emerging from domain-general reward-based learning, yet, left open to address the timescales required to acquire them. Here, we investigate how quickly agents learn to rely on social information under different levels of environmental uncertainty, using boundedly rational reinforcement learning dynamics. We find that learning together not only yields better results under environmental uncertainty but can also significantly speed up learning, provided that agents have a reliable internal model of the environment. When such an internal model is not available, learning together still yields better results, but the learning speed advantage is lost. However, even model-free learning does not take significantly longer than independent learning to achieve most of the final reward that an independently learning agent achieves. Our findings suggest a concrete design principle for hybrid collective intelligence: the quality of the internal world models determines whether collaborative learning accelerates or decelerates.

Keywords Social learning strategies • Internal timescales • Environmental uncertainty • Reinforcement learning • Nonlinear dynamics

1 Introduction

The increasing integration of advanced artificial agents into everyday human activities presents a fundamental challenge to our understanding of collective behavior [Rahwan et al., 2019, Brinkmann et al., 2023]. For example, when deciding on a restaurant to eat at, a hotel to stay in, or a product to buy, humans often consult online reviews and ratings from other users. These reviews nowadays are frequently summarized, generated, and moderated by AI systems on a much faster timescale. Likewise, they are also used by advanced AI agents to make and execute decisions on behalf of their human users, creating complex feedback loops. As humans and AI begin to learn, cooperate, and make decisions in shared environments, we must re-examine the principles that underpin sociality itself. How do adaptive social learning strategies emerge and evolve when humans and AI agents must learn from one another to navigate uncertain worlds? In this article, we show that these rules governing when and from whom to learn emerge endogenously through domain-general reward learning, and we quantify for the first time how quickly this emergence occurs under different environmental and cognitive conditions.

To understand how populations adapt, the field of cultural evolution offers a robust theoretical framework for modeling the transmission and refinement of behaviors [Boyd and Richerson, 1988][Rogers, 1988]. This approach is invaluable for studying the evolution of sociality, as it provides a mathematical basis for analyzing how individuals choose between learning from their own experience and learning from others using various social learning strategies [Acerbi et al., 2022]. These are assumed to be fixed heuristics tailored to particular situations that individuals employ in learning from others. At least 25 distinct social learning strategies, such as 'copy the majority' or 'copy the successful', have been documented in various studies [Laland, 2004, Rendell et al., 2010, Morgan et al., 2011, Kendal et al., 2018]. This approach has transformed the study of cultural evolution from a descriptive analogy to a rigorous field of theoretical and empirical research [Deffner et al., 2024]. However, this fixed heuristics account treats social learning and individual (or 'asocial') learning as conceptually distinct, alternative processes. As such, it cannot explain the origins of these strategies, i.e., the cognitive and environmental factors that give rise to them [Heyes, 2016, Tump et al., 2024].

Rational decision-making models address this limitation by proposing that social learning arises from a single inferential process [Perreault et al., 2012]. Instead of applying a fixed heuristic, individuals use Bayesian inference to rationally update their beliefs about the world, combining social and nonsocial cues [Pérez-Escudero and de Polavieja, 2011, Arganda et al., 2012, Bikhchandani et al., 1992][Banerjee, 1992]. This allows agents to rationally weigh social information according to its reliability and predictive value, leading to social learning strategies that emerge from fundamental information asymmetries [Arganda et al., 2012, Mann, 2018, 2020]. For example, individuals place a heavy weight on social cues when the environment changes slowly or when its state cannot be well predicted using nonsocial cues [Perreault et al., 2012, Mann, 2020]. Rational models of Bayesian inference also predicted people's social learning behaviour better than simpler fixed heuristics [Taylor-Davies et al., 2025]. However, these rational models still have a significant limitation. They excel at explaining an optimal social learning strategy at equilibrium, but do not address the process of how the agent initially acquired it. They therefore cannot answer questions about the timescale for acquiring social learning strategies or the necessary conditions for their emergence. Thus, they cannot account for the adaptability of social learning to changing environmental contexts, as recent experimental findings show [Efferson et al., 2008, Toelch et al., 2014, Deffner et al., 2020, Toyokawa and Gaissmaier, 2022].

To address this gap, an emerging consensus points to the underlying mechanism of domain-general reward learning [Heyes, 2012, Heyes and Pearce, 2015]. This view posits that seemingly complex social learning strategies are not fixed rules but emerge from a more fundamental process: associating social cues with expected rewards. Indeed, evidence from neuroscience [Behrens et al., 2008, Olsson et al., 2020, Zhang and Gläscher, 2020], computational cognitive science [Suganuma et al., 2025, Najar et al., 2020, Bergerot et al., 2024, Witt et al., 2024, Schultner et al., 2025a][Danwitz and von Helversen, 2025], and machine learning research [Borsa et al., 2019, Ndousse et al., 2021, Ha and Jeong, 2023] supports the view that reward learning is central to the emergence of social learning strategies. This body of work strongly suggests that the sophisticated heuristics of social learning are not pre-programmed but are themselves learned and refined through reward. The key advantage of this approach is its adaptation timescale. Learning from experience can happen flexibly within an individual's lifetime. By contrast, the fixed-heuristic approach typically models the evolutionary success of a limited set of fixed heuristics under natural selection [Rendell et al., 2010, Aoki and Feldman, 2014, Turner et al., 2023, Sigalou and Mann, 2023]. As learning operates on a much faster timescale than natural selection, human social learning strategies are supposedly more flexible than is commonly assumed [Schultner et al., 2025b].

However, a systematic investigation of the adaptation timescales remains elusive. It is well known that environmental uncertainty promotes social learning because social cues provide valuable information that complements ambiguous environmental signals [Perreault et al., 2012, Mann, 2018][Wu et al., 2025]. Concerning learning timescales, Suganuma et al. [2025] recently found differences in the internal learning speed of different social reinforcement learning algorithms, such as value shaping and decision biasing, that rely invariably on social information. But how does environmental uncertainty affect the rate of acquiring social learning strategies compared to independent learning from a pure domain-general reward learning framework, where agents can flexibly choose whether or not to integrate social information [Schultner et al., 2025a]? Evidence from machine learning research suggests that the mere presence of an expert teacher agent not only consistently improves the final policy of a learning agent but also accelerates the convergence of learning [Borsa et al., 2019]. However, what if there is no expert teacher to begin with? A reinforcement learning agent that incorporates social cues must learn more complex strategies than one that relies solely on environmental cues. When two reinforcement learning agents learn simultaneously from environmental signals and from each other, achieving higher performance at convergence may come at the cost of slower learning.

In this paper, we demonstrate that this intuition is largely wrong. To do so, we investigate how environmental uncertainty affects both the final performance and the learning speed of emergent social learning strategies within a reward-based learning framework. Specifically, we ask: (1) Under what environmental conditions do learned social learning strategies outperform independent learning? (2) What types of social learning strategies emerge from reward learning under

different levels of uncertainty? (3) How quickly can agents acquire effective strategies, and does learning together accelerate or hinder this process?

To address these questions, we develop a mathematical framework based on partially observable stochastic games, where agents must make decisions based on noisy observations of an environmental state. We operationalize environmental uncertainty in two dimensions: i) observational uncertainty and ii) state uncertainty. We compare single-agent scenarios, in which individuals learn only from environmental feedback, with two-agent scenarios, in which a second agent can observe and potentially learn from the first agent's choices. We examine both model-based learning dynamics (in which agents behave as if they possess an internal model of the environment) and model-free dynamics (in which agents learn purely through trial-and-error experience).

Our analysis reveals four key findings. First, we demonstrate that social learning strategies consistently emerge and outperform independent learning across a wide range of environmental uncertainties, with benefits reaching up to 14% in reward performance. Second, we show that the interplay between observational and environmental state uncertainty gives rise to a rich diversity of adaptive social learning strategies—from simple 'copy-when-uncertain' heuristics to complex conditional strategies that depend on which environmental state is more likely. Third, we find that the transitions between these learned social learning strategies are critical, marked by a significant slowdown in learning at these points. Fourth, we find that away from these critical points, agents with internal environmental models learn approximately twice as fast when paired together as when learning alone, whereas agents without such models show the opposite pattern. Yet remarkably, the time required for learning together to match almost the final performance of independent learners remains comparable, even without internal models.

These findings have important implications for understanding collective behavior in hybrid human-AI populations. As AI systems increasingly mediate human decision-making and learning through content curation, recommendation algorithms, and automated decision support, our results suggest that the speed and efficiency of collective adaptation will depend critically on whether these systems can build reliable internal models of their environments. More broadly, our work demonstrates that the sophisticated social learning strategies observed across human and animal societies need not be innate or evolved over generations—they can emerge rapidly within individual lifetimes through domain-general reward-learning mechanisms.

2 Methods

2.1 Mathematical framework

We formulate our model in the general framework of partially observable stochastic games $G = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{O}, T, R, O \rangle$. There are $\mathfrak{N} \in \mathbb{N}$ agents. The choice environment consists of $\mathfrak{S} \in \mathbb{N}$ states, $\mathcal{S} = (\mathrm{S}_1, \dots, \mathrm{S}_{\mathfrak{S}})$. In each state s, each agent $i \in \mathcal{N} = (1, \dots, N)$ has $\mathfrak{A} \in \mathbb{N}$ available actions $\mathcal{A}^i = (\mathrm{A}^i_1, \dots, \mathrm{A}^i_{\mathfrak{A}})$ to choose from. We denote the joint action set by $\mathcal{A} = \bigotimes_i \mathcal{A}^i$, where $\bigotimes_i$ is a cross-product sign. A joint action by all agents is denoted by $a = (a^1, \dots, a^N) \in \mathcal{A}$, and $a^{-i} = (a^1, \dots, a^{i-1}, a^{i+1}, \dots, a^N)$ denotes the joint action by all agents except agent i. Agents choose their actions simultaneously.

We use the same number of actions, $\mathfrak{A}$, and observations, $\mathfrak{O}$, (see below) across all agents and states for notational and computational convenience. Doing so is computationally convenient, as it allows us to formulate the environmental state transition, reward, and observation functions, T, R, and O, as tensors, thereby enabling accelerated linear algebra (XLA) operations [1].

The transition tensor $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ determines the probability of a state change. $T^{s,a,\acute{s}}$ is the probability to transition from the current state s under the joint action a to the next state $\acute{s}$.

The reward tensor $R : \mathcal{N} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ encodes the immediate rewards the agents receive. $R^{i,s,a,\acute{s}}$ is the reward of agent i when the environment transitions from state s, under the joint action a, to state $\acute{s}$.

Instead of observing the states $s \in \mathcal{S}$ directly, each agent i observes one of $\mathfrak{O} \in \mathbb{N}$ observations $\mathcal{O}^i = (\mathrm{O}^i_1, \dots, \mathrm{O}^i_{\mathfrak{O}})$ according to its observation tensor $O^i : \mathcal{S} \times \mathcal{O}^i \to [0, 1]$. O^{i,s,o^i} is the probability that agent i observes observation $o^i \in \mathcal{O}^i$ when the environment is in state s. The joint observation set is $\mathcal{O} = \bigotimes_i \mathcal{O}^i$ and $O = \bigotimes_i O^i : \mathcal{N} \times \mathcal{S} \times \mathcal{O} \to [0, 1]$ is the joint observation tensor.

Agents select actions according to their strategies $X^i : \mathcal{O}^i \times \mathcal{A}^i \to [0, 1]$ (also known as policies). X^{i,o^i,a^i} is the probability that agent i selects action a^i when it observes observation o^i. The joint strategy tensor is $X = \bigotimes_i X^i : \mathcal{N} \times \mathcal{O} \times \mathcal{A} \to [0, 1]$.

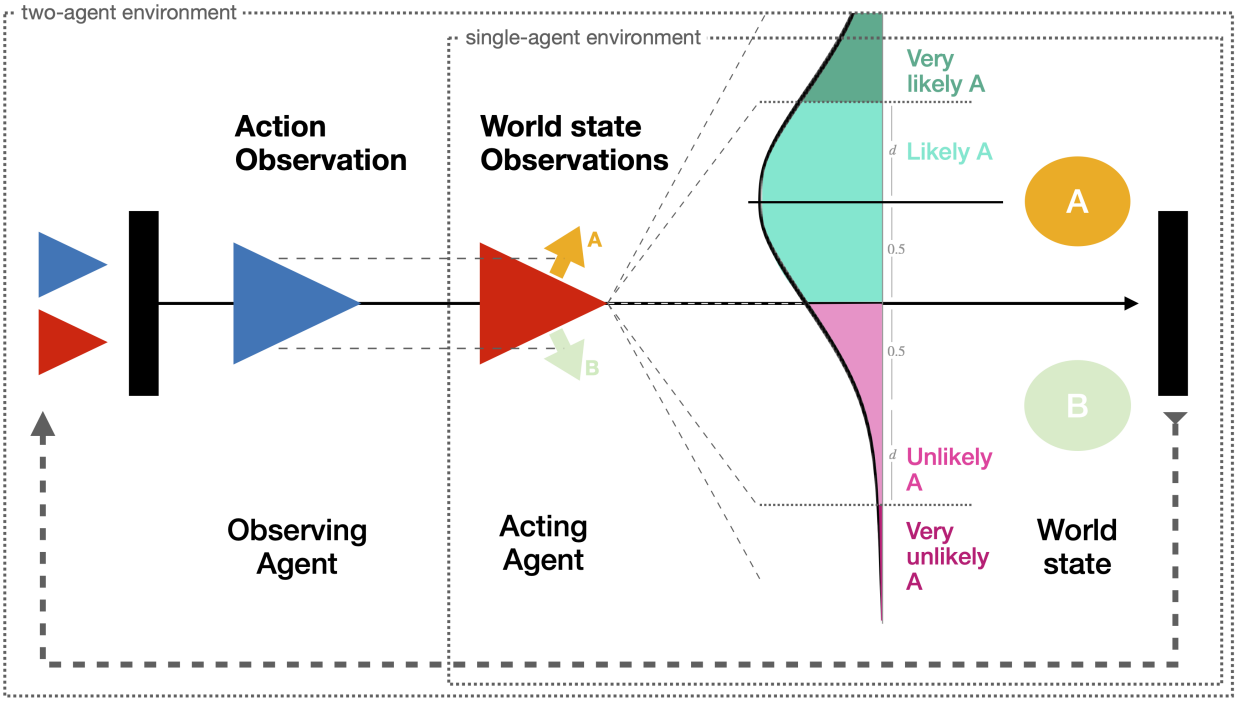

Figure 1: **Choice environments.** In the single-agent environment, the agent observes a noisy signal about the state of nature, A or B. If the agent acts in accordance with the state of nature, it receives a high reward; otherwise, it gets a low reward. The two-agent environment builds upon the single-agent environment. A second agent observes the choice taken by the first agent. Thus, the second agent can learn to condition its own action choices on the other agent's choices. Which agent is first is decided by a coin flip after each trial.

2.2 Choice environment

We consider two choice environments: a single-agent environment and a two-agent environment (Figure 1). In the single-agent environment, the agent observes a noisy signal about the state of nature. If the agent acts in accordance with the state of nature, it receives a high reward; otherwise, it gets a low reward. The two-agent environment builds upon the single-agent environment. A second agent observes the choice taken by the first agent. Thus, the second agent can learn to condition its own action choices on the other agent's choices.

Single-agent environment. Specifically, the single-agent environment consists of two states of nature, $\mathcal{S}_{\text{single}} = \{\text{A}, \text{B}\}$, and the agent can choose between two actions $\mathcal{A}^1 = \{\text{A}, \text{B}\}$. If the agent chooses action A (B), when the environment is in state A (B), the agent receives a reward $R^{1,\text{A},\text{A},\acute{s}} = R^{1,\text{B},\text{B},\acute{s}} = 1.0$, regardless of the next state $\acute{s}$. Otherwise, the agent receives a reward $R^{1,\text{A},\text{B},\acute{s}} = R^{1,\text{B},\text{A},\acute{s}} = -1.0$. However, the agent cannot directly observe the state of nature.

Observational uncertainty. The agent observes one of four possible signals, which we denote by $\mathcal{O}^1 = \{\text{A}, \text{a}, \text{b}, \text{B}\}$. Here, the signals A and a denote that nature is more likely to be in state A, whereas signals B and b denote that nature is more likely to be in state B. Capital letters indicate more certainty than lower-case letters regarding the respective signal-state pairs. Thus, observing A indicates that nature is in A is *very likely*, whereas observing a means that nature is only *likely* in A. For state B, the analogous applies. Note that the agent has no capacity to decipher the meaning of these letters. It only learns via observation-reward associations.

We model these observation probabilities by discretizing a normal distribution $\mathcal{N}_{\mu,\sigma}$, which is centered around the true state of nature ($\mu = 0$). The other state is imagined to reside one unit length apart. Thus, if the true state of nature is A (B),

- the agent observes signal A (B) with probability $O^{1,\text{A},\text{A}} = O^{1,\text{B},\text{B}} = \int_{-\infty}^{-d} \mathcal{N}_{0,\sigma}(x)\, dx$;
- it observes signal a (b) with probability $O^{1,\text{A},\text{a}} = O^{1,\text{B},\text{b}} = \int_{-d}^{0.5} \mathcal{N}_{0,\sigma}(x)\, dx$;

[1] https://openxla.org/

- it observes signal b (a) with probability $O^{1,\mathrm{A},\mathrm{b}} = O^{1,\mathrm{B},\mathrm{a}} = \int_{0.5}^{1+d} \mathcal{N}_{0,\sigma}(x)\,dx$; and last
- it observes signal B (A) with probability $O^{1,\mathrm{A},\mathrm{B}} = O^{1,\mathrm{B},\mathrm{A}} = \int_{1+d}^{\infty} \mathcal{N}_{0,\sigma}(x)\,dx$,

where $d > 0$ is the distance from the center of nature's true state at which the agent believes to observe the state *very* likely. Without loss of generality, we set $d = 0.5$ throughout this paper. We use the standard deviation of the normal distribution, σ, to represent the observational uncertainty of our model.

State uncertainty. We model the next state of the world as independent of the current state and actions. It is determined by a parameter p_{A}, denoting the probability of transitions to nature state A, i.e., $T^{s,a,\mathrm{A}} = p_{\mathrm{A}}$ and $T^{s,a,\mathrm{B}} = 1 - p_{\mathrm{A}}$. We will investigate the range $p_{\mathrm{A}} > 0.5$. Thus, we denote the state uncertainty of our model by the probability that nature is in state B, $p_{\mathrm{B}} = 1 - p_{\mathrm{A}}$. This way, when $p_{\mathrm{B}} = 1 - p_{\mathrm{A}} = 0.5$, the environment is maximally uncertain, as both states occur equally likely. By contrast, when p_{B} approaches 0, nature becomes more predictable, and state uncertainty decreases. For example, taking the bus might be the best way to get to work on 9 out of 10 days. But you could also walk if you receive a strong enough signal that the bus will be severely delayed, e.g., when you see snow and ice on the streets. Note that most other works, e.g., [Mann, 2018, 2020, Sigalou and Mann, 2023], have not really explored this kind of uncertainty and always assumed $p_A = p_B = 0.5$ for convenience.

These two dimensions map onto distinct features of empirical social learning environments, and their interplay determines when observing others is most valuable. Consider a traveler exploring a new city whose culinary scene spans Chinese, Greek, Italian, Mexican, and Vietnamese cuisine, among others. Each day, the traveler narrows the choice to two restaurants. The hidden state of the world captures which of the two is the genuine better fit for the traveler. This is determined by the traveler's personal tastes and the intrinsic quality of each restaurant. State uncertainty (p_B) reflects how predictable this fit is: when the traveler has a strong preference for one cuisine over the other, one option is reliably the better choice (low p_B); when the two cuisines are equally appealing, the better choice is unpredictable ($p_B \approx 0.5$). Observational uncertainty (σ) captures how clearly available cues (e.g., aromas drifting from the kitchen, the displayed menu, the restaurant's atmosphere) signal which option is the better fit on this day. Low observational uncertainty means these cues are strong and diagnostic. High observational uncertainty means they are weak or ambiguous. The two dimensions have distinct consequences for social learning: when cues are ambiguous (high σ), a fellow traveler's restaurant choice provides a valuable independent signal (assuming similar tasts). When one option almost always fits better (low p_B), even imperfect private cues tend to identify the right option, reducing the marginal value of copying others.

Two-agent environment. The two-agent environment adds a second agent to the uncertain decision problem described above. The natural state, $\mathcal{S}_{\text{single}} = \{\mathrm{A}, \mathrm{B}\}$, is complemented by a social component that describes the sequence of agents' decision points. Agents choose sequentially. The first-acting agent can rely only on the natural observation signal to decide on an action. The subsequent agent observes this action along with the natural observation signal to decide on its action. After the second choice is made, both agents receive their rewards, and the process restarts, with the agents' order determined at random. The social state set describes this agent sequence, $\mathcal{S}_{\text{social}} = \{12., .1\underline{\mathrm{A}}, .1\underline{\mathrm{B}}, 21., .2\underline{\mathrm{A}}, .2\underline{\mathrm{B}}\}$. The first letter of each state identifier denotes the non-acting agent $i \in \mathcal{N} = \{1, 2\}$. The second letter describes the acting agent. The last letter indicates the action choice of the first-acting agent $a \in \{\underline{\mathrm{A}}, \underline{\mathrm{B}}\}$. If there is no such action, or no non-acting agent, a '.' signifies an empty spot. Thus, the combined social and natural state set of the two-agent environment consists of 12 states,

$$\mathcal{S}_{\text{two}} = \mathcal{S}_{\text{single}} \times \mathcal{S}_{\text{social}} = \{\mathrm{A}12., \mathrm{A}.1\underline{\mathrm{A}}, \mathrm{A}.1\underline{\mathrm{B}}, \mathrm{A}21., \mathrm{A}.2\underline{\mathrm{A}}, \mathrm{A}.2\underline{\mathrm{B}}, \mathrm{B}12., \mathrm{B}.1\underline{\mathrm{A}}, \mathrm{B}.1\underline{\mathrm{B}}, \mathrm{B}21., \mathrm{B}.2\underline{\mathrm{A}}, \mathrm{B}.2\underline{\mathrm{B}}\}.$$

The observation set of the two-agent environment combines the noisy observations of the natural state with the observation about the first-acting agent plus a dummy observation, '..', for the non-acting agent,

$$\mathcal{O}^i_{\text{two}} = \{\mathtt{A\underline{.}},\ \mathtt{a\underline{.}},\ \mathtt{b\underline{.}},\ \mathtt{B\underline{.}},\ \mathtt{A\underline{A}},\ \mathtt{a\underline{A}},\ \mathtt{b\underline{A}},\ \mathtt{B\underline{A}},\ \mathtt{A\underline{B}},\ \mathtt{a\underline{B}},\ \mathtt{b\underline{B}},\ \mathtt{B\underline{B}},\ \mathtt{..}\} \quad \text{for } i = 1, 2,$$

where the first letter of each observation identifier denotes the natural observational signal, and the second letter represents the action taken by the first-acting agent. If there is no such action because it's the first-acting agent's turn, a '.' signifies an empty spot.

In the two-agent environment, each round consists of two sequential actions before rewards are distributed, so each agent receives reward once per round rather than once per action. To make average rewards comparable across the two environments, we multiply rewards in the two-agent environment by 2 to compensate for the sparser feedback. This rescaling encodes the assumption that social observation is temporally free: the time the second agent spends waiting within a round carries no real-time cost. This is the natural baseline in many real-world social learning contexts, where

social cues are largely ubiquitous and available passively without dedicated observation effort. Under this convention, iteration counts of the learning rule measure cognitive convergence cost rather than individual sample cost.

2.3 Learning dynamics

We consider deterministic reinforcement learning dynamics [Bloembergen et al., 2015, Barfuss and Mann, 2022], which provide a transparent and analytically tractable framework for studying the collective behavior of individual learners [Barfuss et al., 2025]. For example, they allow for a straightforward definition of convergence, enabling the analysis of learning timescales. With a formalism similar to that of evolutionary game theory [Börgers and Sarin, 1997, Tuyls et al., 2003, Sato and Crutchfield, 2003][Bernasconi et al., 2025], our reinforcement learning model boils down to the exponential replicator dynamics in discrete time [Hofbauer and Sigmund, 1998],

$$X^a_{t+1} \propto X^a_t \exp(\alpha F^a_{X_t}).$$

The probability of choosing action a at time step t, X^a_t, is updated proportionally to its exponential fitness $F^a_{X_t}$ under the current strategy profile X_t, scaled by the learning rate α. Without loss of generality, we set the learning rate $\alpha = 0.25$ for all agents and environments throughout this article .In the Supplementary Information, we demonstrate that our main results are robust to varying $\alpha \in \{0.125, 0.25, 0.5\}$. As one would expect, the smaller α, the longer the convergence time becomes.. This dynamic captures how agents adjust their action preferences based on the rewards they receive, leading to the evolution of their strategies over time. Action probabilities of high-performing actions increase, whereas those of underperforming actions decrease. High-performing actions are *replicated* more than underperforming ones.

As we seek to investigate the learning timescales across different levels of environmental uncertainty, we consider two types of agents: model-based and model-free reinforcement learning agents. Model-based agents operate as if they possess an internal model of the environment as they perceive it, allowing them to plan their actions by simulating future observations and rewards. In contrast, model-free agents learn to associate actions only through experienced rewards via trial-and-error interactions with the environment [Sutton and Barto, 2018]. Whether an agent has an internal model is likely to influence how quickly it can acquire effective (social learning) strategies.

As our deterministic dynamics operate on a higher level of cognitive abstraction than typical reinforcement learning algorithms [Barfuss, 2022], we do not explicitly model the internal processes of planning or value function estimation. Instead, we capture the essence of model-based and model-free learning by directly computing fitness values that drive the learning updates. Our two dynamics should, therefore, be seen as idealized representations of model-based and model-free learning processes, respectively, where the model-based dynamics operate as if they have a perfect internal model of the current environment they perceive.

Model-based learning dynamics. The update of the agents' strategies of our model-based learning dynamics yields,

$$X^{i,o^i,a^i}_{t+1} = \frac{X^{i,o^i,a^i}_t \exp\left(\alpha^i R^{i,o^i,a^i}_{X_t}\right)}{\sum_{b\in\mathcal{A}^i} X^{i,o^i,b}_t \exp\left(\alpha^i R^{i,o^i,b}_{X_t}\right)},$$

where α^i denotes agent i's learning rate and $R^{i,o^i,a^i}_{X_t}$ is the average reward agent i receives when it observes observation o^i and takes action a^i, given that all agents behave according to the current strategy X_t. Thus, the 'fitness' of action a^i under observation o^i is simply the expected reward for taking that action in that observational context, even if that context is rarely observed or that action is seldom taken. Therefore, we call this update rule *model-based*.

To perform the learning update, we need to compute the average rewards R^{i,o^i,a^i}_X. However, the environment's rewards $R^{i,s,a,\acute{s}}$ depend on the states s, not the observations o^i. To obtain the average observation-action rewards R^{i,o^i,a^i}_X, we require a mapping from observations to states to account for the states s underlying a given observation o^i. The observation tensor O^{i,s,o^i} accounts for the observations o^i following from a state s. Following Barfuss and Mann [2022], we can invert the observation tensor into a belief tensor using Bayes' rule,

$$B^{i,o^i,s}_X = \frac{O^{i,s,o^i} P^s_X}{\sum_s O^{i,s,o^i} P^s_X},$$

where P^s_X is the stationary state distribution given the strategy X. Thus, $B^{i,o^i,s}_X$ is the belief of agent i, that the environment is in state s, given it observes observation o^i. The stationary state distribution, P^s_X, is the left eigenvector of the strategy-average transition matrix $T^{s,\acute{s}}_X$. It can be obtained via

$$T_X^{s,\acute{s}} = \sum_{a^j \in \mathcal{A}^j} \prod_{j \in \mathcal{N}} Y_X^{j,s,a^j} T^{s,a,\acute{s}},$$

where Y_X^{j,s,a^j} are the effective state-action strategies. They can be obtained by averaging out the observations,

$$Y_X^{j,s,a^j} = \sum_{o^j \in \mathcal{O}^j} O^{j,s,o^j} X^{j,o^j,a^j}.$$

With Y_X^{j,s,a^j} we can obtain the effective transition matrix, $T^{s,\acute{s}}$, the stationary state distribution as its left eigentvector, P_X^s, and from that, the belief tensor $B^{i,o^i,s}$. With $B^{i,o^i,s}$, we can finally compute the strategy-average observation-action rewards R_X^{i,o^i,a^i}.

Whenever agent i observes observation o^i, the environment is state s with probability $B^{i,o^i,s}$. In s, all other agents $j \neq i$ behave according to Y_X^{j,s,a^j}. The environment transitions to state $\acute{s}$ with probability $T^{s,a,\acute{s}}$. And agent i receives reward $R^{i,s,a,\acute{s}}$. Thus, the strategy-average observation-action reward for action a^i under observation o^i reads,

$$R_X^{i,o^i,a^i} = \sum_{s \in \mathcal{S}} \sum_{a^j \in \mathcal{A}^j} \sum_{\acute{s} \in \mathcal{S}} \sum_{j \neq i} B_X^{i,o^i,s} Y_X^{j,s,a^j} T^{s,a,\acute{s}} R^{i,s,a,\acute{s}}.$$

Model-free learning dynamics. In addition to the model-based learning dynamics described above, we also consider model-free learning dynamics. In model-free learning, agents do not maintain an internal model of the environment's transition, reward, and observation tensors. Hence, they cannot use the average rewards R_X^{i,o^i,a^i} to update their strategies. They have to be taken proportionally to the agents' experience. The agents' experience is captured by multiplying the average rewards with the stationary observation distribution P_X^{i,o^i}, and the strategy X^{i,o^i,a^i}. Thus, the reinforcement signal for an action a^i under an observation o^i is proportional to the average reward of that action times the frequency with which the action is taken under the observation o^i times the frequency of observing o^i. Therefore, the model-free learning dynamics read,

$$X_{t+1}^{i,o^i,a^i} = \frac{X_t^{i,o^i,a^i} \exp\left(\alpha^i P_{X_t}^{i,o^i} X_t^{i,o^i,a^i} R_{X_t}^{i,o^i,a^i}\right)}{\sum_{b \in \mathcal{A}^i} X_t^{i,o^i,b} \exp\left(\alpha^i P_{X_t}^{i,o^i} X_t^{i,o^i,b} R_{X_t}^{i,o^i,b}\right)}.$$

The stationary observation distribution P_X^{i,o^i} can be obtained by averaging the observation tensor O^{i,s,o^i} over the stationary state distribution P_X^s,

$$P_X^{i,o^i} = \sum_{s \in \mathcal{S}} O^{i,s,o^i} P_X^s.$$

Relation to standard reinforcement learning algorithms. The terminology model-based and model-free spans a wide algorithmic space, including planning methods such as Dyna and temporal-difference methods such as Q-learning [Sutton and Barto, 2018], and our dynamics should not be equated with any specific algorithm within these families. The property our dynamics *do* capture is the core informational distinction: whether the agent uses an internal model of the environment's transition, reward, and observation structure when computing its update signal. Our model-based dynamics do so explicitly, constructing Bayesian beliefs about the hidden state (see below) and computing expected rewards over the full joint dynamics. Our model-free dynamics instead weight updates only by empirically determined observation visit frequencies P_X^{i,o^i}, as an agent without access to the underlying state structure would. The properties our dynamics do *not* capture are algorithm-specific implementation details. The dynamics are deterministic, computing exact expectations over the stationary distribution rather than sampling individual transitions; this corresponds to the mean-field or population limit of stochastic RL algorithms [Börgers and Sarin, 1997, Tuyls et al., 2003]. Neither dynamics explicitly maintain value functions, perform multi-step planning, or use temporal-difference bootstrapping. This idealization isolates the informational mechanism from implementation details, providing a clean theoretical reference point for what each learning paradigm can achieve in this environment.

3 Results

We will first investigate how the final performance differences between individual and emergent social learning depend on environmental uncertainty and the learning model. Second, we analyze the learned strategies that underlie these

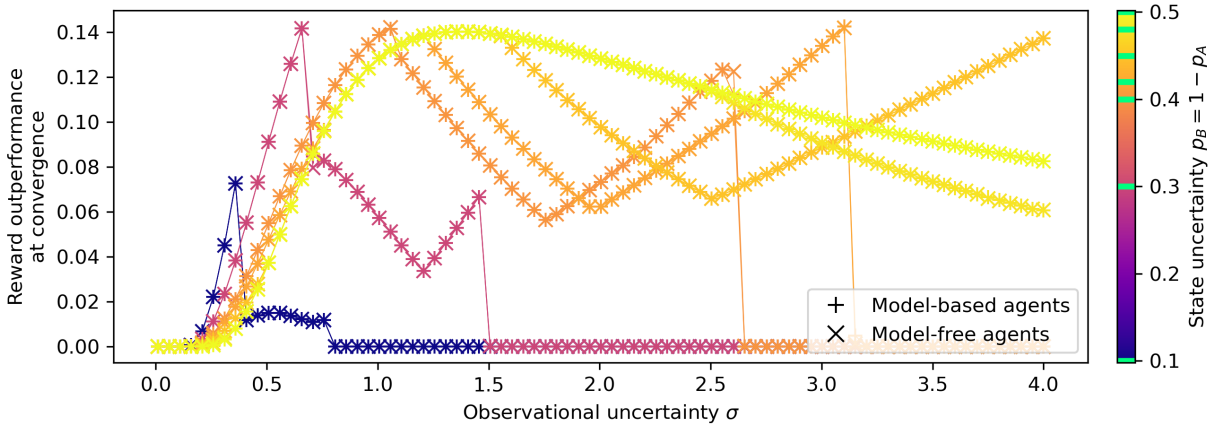

Figure 2: **Final Reward Outperformance** versus observational uncertainty σ for different state uncertainties p_{B} (shown in color, from dark blue at $p_{\mathrm{B}} = 0.1$ to yellow at $p_{\mathrm{B}} = 0.5$) and both learning dynamics: model-free (x) and model-based (+). Outperformance is defined as the relative reward gain of two-agent over single-agent learning, $(R_{\text{social}} - R_{\text{single}})/R_{\text{single}}$; positive values indicate that social learning yields higher rewards. Two-agent learning consistently outperforms independent learning across all tested levels of environmental uncertainty, with the largest benefit at intermediate observational uncertainty.

performance differences. Lastly, we study the learning timescales of learning together versus learning alone. We define a learning process as converged when the change in the strategy space, composed of the action probabilities across all agents, observations, and actions, falls below a threshold $\epsilon = 10^{-6}$. In the Supplementary Information we show that our main results are robust to the choice of convergence threshold by varying $\epsilon \in \{10^{-5}, 10^{-6}, 10^{-7}\}$. As one would expect, the smaller ϵ, the longer the convergence time becomes. The shape of the final reward outperformance curves is unaffected by this choice..

3.1 Final performance differences

With more environmental uncertainty, even optimally learned strategies yield lower rewards because the environment is less predictable. Thus, we are not measuring reward performance in absolute terms. Instead, we measure the *outperformance* of learning social learning strategies relative to independent learning strategies. We define outperformance as the relative difference in average rewards between agents that can learn from social and environmental cues, R_{social} versus agents that can learn only from environmental cues, R_{single}, i.e., the reward outperformance amounts to $(R_{\text{social}} - R_{\text{single}})/R_{\text{single}}$. Positive outperformance indicates that learning social learning strategies yields higher average rewards than independent learning strategies, whereas negative outperformance indicates the opposite.

Figure 2 establishes that learning social learning strategies (both model-based and model-free) consistently outperform independent learning strategies across a wide range of environmental uncertainties. Model-based and model-free learning achieve identical final reward outperformance levels, indicating that their differences do not affect their final outcomes. They achieve an outperformance level of up to 14% across most levels of observational and state uncertainty. Only for very low state uncertainty $p_B \leq 0.1$, the maximal outperformance drops below 10%. When state uncertainty is very low, there is little benefit from social learning, as the environment is already quite predictable from the agent's own observations. Importantly, they never underperform compared to independent learning strategies.

The yellow line in Figure 2 depicts the reward outperformance curve at convergence when state uncertainty is maximal ($p_B = 0.5$). At very low observational uncertainty, social learning provides minimal advantage because the environment is already quite predictable from individual observations alone. The outperformance rises sharply between $\sigma \approx 0.3$ and peaks at intermediate observational uncertainty levels around $\sigma \approx 1.4$. Beyond this peak, outperformance declines gradually as observational uncertainty increases further and social cues become increasingly unreliable. Thus, an intermediate level of observational uncertainty provides the optimal balance for social learning to outperform independent learning strategies.

When state uncertainty decreases ($p_B < 0.5$), the final reward outperformance curve as a function of observational uncertainty σ reveals its characteristic shape (e.g., for $p_B = 0.4$). This shape persists across all levels of state uncertainty, but becomes increasingly compressed as state uncertainty p_B decreases. It is characterized by an initial sharp rise in outperformance at low observational uncertainty levels, reaching a maximum at intermediate levels, followed by a

gradual decline, then a secondary increase, before eventually dropping to zero at high observational uncertainty. This compression indicates that as the environment becomes more predictable from individual observations, the range of observational uncertainty levels for which learning together provides a significant advantage narrows. As the state of nature becomes more certain, short-term observations about it become less relevant.

Interestingly, the outperformance shape for maximal state uncertainty ($p_B = 0.5$, yellow line in Figure 2) serves as a reference point for understanding how reduced state uncertainty affects learned social learning benefits. For moderately reduced state uncertainty (with p_B between 0.48 and 0.42), the outperformance curves closely follow the reference line at low observational uncertainty levels. The curves begin to diverge only as observational uncertainty increases further. In contrast, when state uncertainty is substantially reduced (such as $p_B \leq 0.4$), the outperformance curves exhibit a more pronounced initial rise at low observational uncertainty than the reference line suggests. Yet after reaching their peak outperformance, these curves decline sharply and continue right below the reference line.

Overall, these results suggest that the final reward outperformance achieved through learned social learning strategies is robust across different learning models and levels of environmental uncertainty. The characteristic shape of the outperformance curve indicates a fundamental relationship between observational and state uncertainty in determining the benefits of learned social learning. To better understand the origins of this shape, we will next analyze the final strategies the agents have acquired at convergence.

3.2 Final strategies

What social learning strategies are learned? Figure 3 reveals that unexpectedly complex strategies emerge from reward learning, especially in environments where the state of nature is not fully uncertain ($p_B < 0.5$).

When the state of nature is maximally uncertain ($p_B = 0.5$), a single type of social learning strategy emerges consistently across intermediate to high levels of observational uncertainty ($\sigma \gtrapprox 0.19$), as shown by the straight vertical gray line in Figure 3 A. In the absence of social information, i.e., when agents observe only the natural environmental signals (A_, a_, b_, B_), they rely exclusively on these private observation signals to guide their decisions, as indicated by the light colors. However, when social information becomes available through observing the first-acting agent's choice, a distinct behavioral pattern emerges. The second-acting agent learns to disregard its own natural observation signal (shown by dark colors) when that signal is weak (i.e., a or b) and contradicts the first-acting agent's choice. At these specific observations (aB and bA), the second-acting agent defers to the first-acting agent's decision rather than following its own uncertain signal. We characterize this strategy as a 'copy-when-uncertain' strategy, in which agents rely on social information only when their private environmental cues are ambiguous. This strategy emerges robustly under both model-based and model-free learning dynamics, suggesting it represents a fundamental adaptive response to environmental uncertainty rather than an artifact of a particular learning mechanism.

At very low observational uncertainty ($\sigma \lessapprox 0.19$), some observational signals occur with extremely low probability, which creates notable differences between model-based and model-free learning dynamics (Figure 3 A). When observational uncertainty is minimal, the strong signals A and B almost never occur. Consequently, agents select actions randomly at these observations, as indicated by the intermediate colors in Figure 3. However, this random behavior has no meaningful impact on final reward outperformance, precisely because such observations are so rare in practice. Model-free learning dynamics are more sensitive to observation probabilities than their model-based counterparts. As a result, the random behavior at strong signals A and B persists across a broader range of observational uncertainty levels for model-free agents compared to model-based agents. Interestingly, model-free agents also randomize their actions at observations aB and bA, the very observations where, at higher uncertainty levels, they learn to completely disregard their natural signals and instead copy the first-acting agent's choice. Model-based agents, by contrast, learn to rely fully on their private observation signals at these low uncertainty levels. Only within a narrow band of observational uncertainty ($0.09 \lessapprox \sigma \lessapprox 0.13$, dashed gray lines in Figure 3) do model-free agents display clear signatures of the 'copy-when-uncertain' strategy, a pattern that holds consistently across all levels of state uncertainty p_B. Despite these differences between learning dynamics, the overall behavioral patterns at very low observational uncertainty remain qualitatively similar across all levels of state uncertainty, suggesting that extreme information scarcity produces comparable learning outcomes regardless of environmental predictability.

When the environment is not fully uncertain ($p_B < 0.5$), a rich diversity of social learning strategies emerges, as illustrated in Figure 3 B. These strategies can be categorized into three primary regimes based on the behavior of the first-acting agent. i) In the first regime, the first-acting agent follows its natural observation signals. ii) The second regime occurs when the first-acting agent begins to ignore weak signals associated with the less-likely environmental state. iii) In the third regime, this behavior extends further, as the agent ignores both weak and strong signals in favor of the more likely state. We now examine the specific social learning strategies that arise within these three regimes across the range of increasing observational uncertainty ($\sigma \gtrapprox 0.13$).

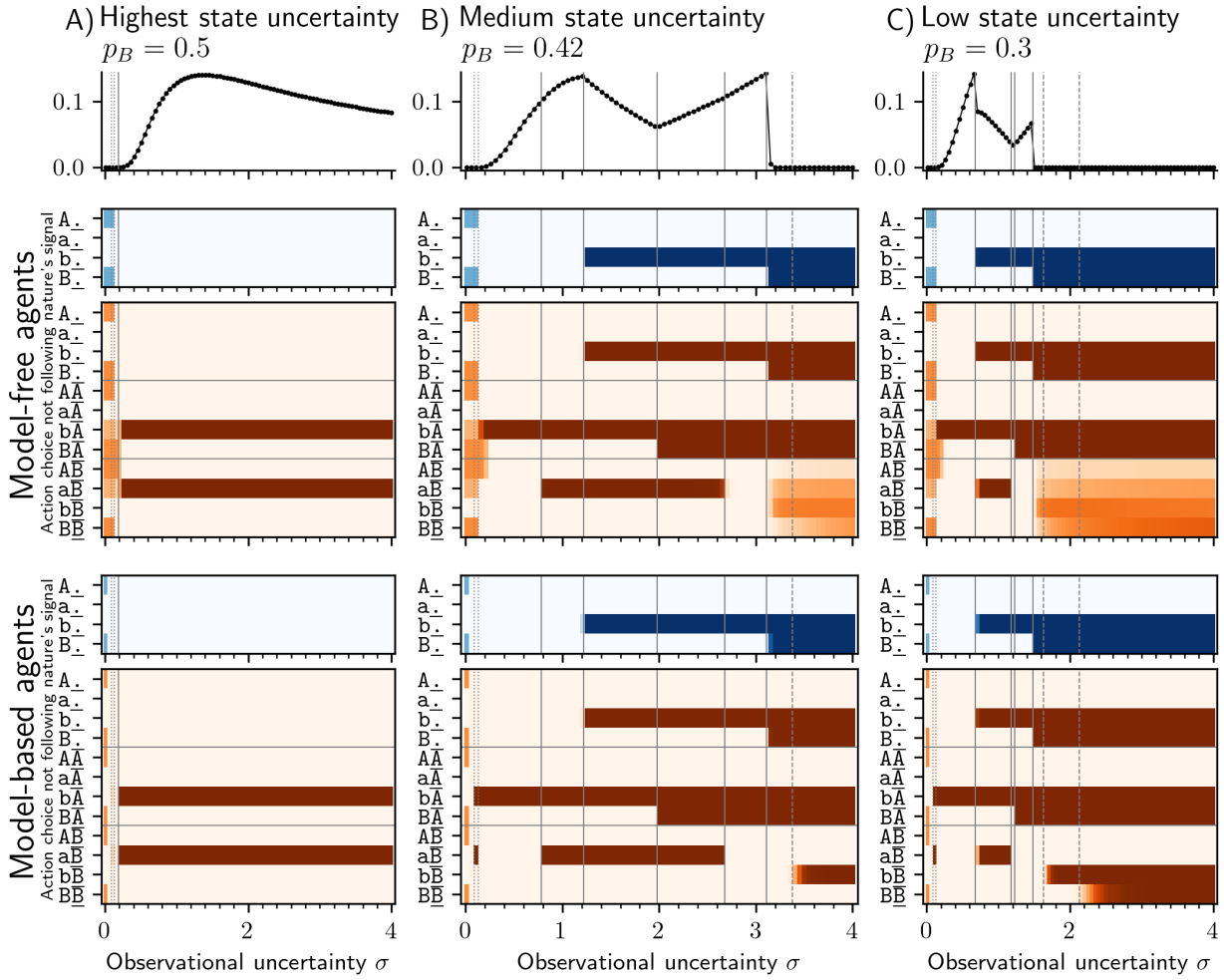

Figure 3: **Learned strategies** at convergence across different levels of observational uncertainty (σ) and state uncertainty $p_B = 0.5$ (A), $p_B = 0.42$ (B), $p_B = 0.3$ (C). The top panels show the reward outperformance curves for reference. The heatmaps display the final learned behaviors across different levels of observational uncertainty (σ) and state uncertainty (p_B). Blue colormaps represent the independent agent in the single-agent environment, while orange colormaps depict behavior in the two-agent environment. The color intensity encodes the final behavioral strategy: light colors indicate that agents follow their natural observation signal, while dark colors indicate that they deviate from it in favor of the alternative action. For example, for observations A_, aB̄ (where the first letter indicates the observation signal), light colors denote choosing action A, whereas dark colors denote choosing action B. Results for the model-based (model-free) agents are averaged over 25 (10) random initial strategies. At high observational uncertainty, agents robustly learn a 'copy-when-uncertain' strategy; richer strategy diversity emerges as state uncertainty decreases.

In the first regime, where observational uncertainty remains relatively low, the second-acting agent develops a strategy we term 'copy-when-uncertain-about-the-less-likely-state.' In this context, the first-acting agent consistently follows its natural observation signals. The second-acting agent, however, learns to selectively ignore its own weak signals when they point toward the less-likely environmental state (indicated by the dark colors at observation bA). In these instances, the agent defers to the first-acting agent's choice. Conversely, when its private signal suggests the more-likely environmental state (as seen at observation aB), the agent continues to rely on its own information to guide its decision. Only as observational uncertainty increases further, specifically reaching a threshold of approximately $\sigma \approx 0.78$ in Figure 3 B, the second-acting agent transitions fully to the 'copy-when-uncertain' strategy. Thus, decreased state uncertainty shifts the emergence of this strategy to higher levels of observational noise. Notably, while this transition marks a clear change in behavioral patterns, it does not significantly impact the shape of the final reward outperformance curve, compared to the fully uncertain environment ($p_B = 0.5$) discussed earlier.

In the second regime, beginning around $\sigma \approx 1.22$ in Figure 3 B, changes in the first-acting agent's strategy cause a decline in reward outperformance. While the second-acting agent initially persists with a 'copy-when-uncertain' approach, it eventually transitions at higher levels of observational uncertainty ($\sigma \approx 1.98$) to a 'follow-the-leader-in-the-more-likely-state-but-copy-when-uncertain-in-the-less-likely-state' strategy. Under this rule, the second agent adopts the first agent's choice when it aligns with the more probable state of the world, effectively ignoring its own private signals, yet only defers to the leader in the less likely state when its own observations are ambiguous. This transition marks a reversal, with reward outperformance beginning to rise again. Eventually, at $\sigma \approx 2.68$ in Figure 3 B, the agent adopts a 'follow-the-leader-in-the-more-likely-state' strategy, where it relies on its own natural observation signals for the less likely state regardless of the leader's action. Interestingly, this final stage is less social than the previous strategy despite increased observational uncertainty. This might be due to the combination of high observational uncertainty and low state uncertainty, which makes the first agent's choice of the less likely state rare and unreliable as a social cue. Note that in the first and second regimes, the learned strategies are consistent across both model-based and model-free learning dynamics.

In the third regime, which begins around $\sigma = 3.11$ in Figure 3 B, the first-acting agent fully ignores its natural observation signal for the less-likely environmental state (indicated by dark colors at B. and b.), at which point the final reward outperformance drops back to zero. During this phase, the second-acting agent continues to follow the leader when they choose the naturally more-likely state (shown by dark colors at bA and BA). When the second-acting agent observes the first-acting agent choosing the less-likely environmental state (B), the model-free learning dynamics randomize their strategy again, since such observations practically never occur. The model-based learning dynamics, by contrast, learn to ignore both their natural observation signal and the first-acting agent's choice (first evident by dark colors at bB from around $\sigma = 3.38$ in Figure 3 B) in order to choose the more-likely environmental state action A. These are the only transitions of the model-based agent that happen gradually rather than sharply. We interpret this behavior as an artifact of our model-based learning dynamics, which can learn from experience they practically never encounter. For this reason, we do not further classify these strategies, though interestingly, a signature of these strategies is also observed in the randomized strategy of the model-free learning dynamics at these same observations.

For lower state uncertainty, these transitions become compressed together, though their topology, i.e., the identity and ordering of transitions, may change (Figure 3 C). For instance, in the first regime, the second-acting agent does not re-adopt the 'copy-when-uncertain' strategy. Instead, when the second regime begins, and the first-acting agent starts disregarding its less likely individual observation of the less likely state (b.), the second-acting agent responds by adopting the 'copy-when-uncertain' strategy. Subsequently, at $\sigma \approx 1.18$ in Figure 3 C, the second-acting agent switches from this 'copy-when-uncertain' strategy to a new 'copy-when-uncertain-about-the-less-likely-state' strategy, before finally transitioning to the 'follow-the-leader-in-the-more-likely-state' strategy at $\sigma \approx 1.23$.

Overall, these results demonstrate that a rich diversity of social learning strategies can emerge from simple reward-based learning dynamics, particularly when environmental state uncertainty is reduced. The specific strategies that agents adopt depend critically on the interplay between observational and state uncertainty. Notably, reward outperformance increases when the second-acting agent learns to disregard its private signal about the less-likely environmental state and instead follows the first-acting agent's choice. Conversely, reward outperformance decreases when the first-acting agent itself begins to ignore its private signals for the less-likely state, thereby reducing the reliability of social information. The information-theoretic basis of this pattern is straightforward: the second-acting agent's observation combines its own noisy private signal with the first-acting agent's choice, which itself integrates that agent's private signal about the same environment. Combining two imperfect signals reduces residual state uncertainty more than either alone, and this benefit is greatest precisely when individual signals are weakest, which is why the copy-when-uncertain strategy emerges under high observational uncertainty. Having characterized these emergent behavioral patterns, we now turn to analyzing the timescales over which agents acquire these social learning strategies.

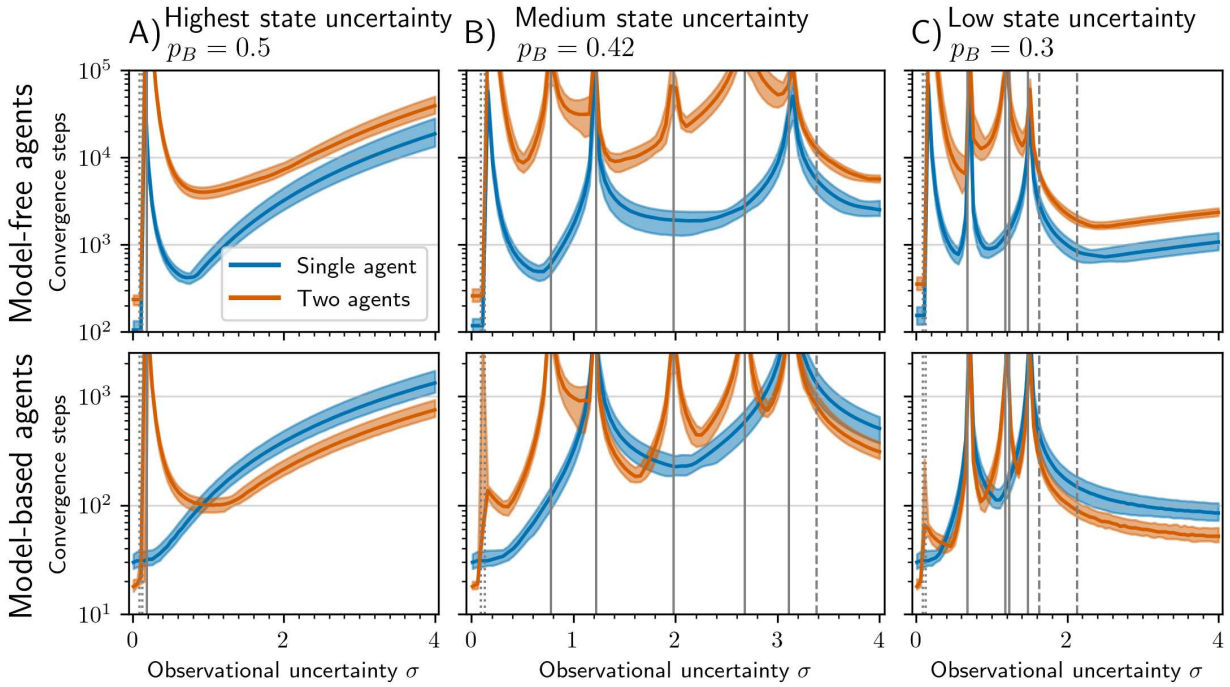

Figure 4: **Learning times** until convergence of the behavioral rules versus observational uncertainty σ for three different state uncertainties $p_B = 0.5$ (A), $p_B = 0.42$ (B), $p_B = 0.3$ (C). The upper panels show model-free learning dynamics, the lower panels show the model-based ones. Blue (vermillion) lines show the single-agent (two-agent) environment. Shaded regions indicate the space between 5- and 95-percentiles from 25 (10) random initial strategies for the model-based (model-free) agents. Convergence times spike at strategy transition points, reflecting a critical slowing down. Uutside these points, model-based agents learn approximately twice as fast in the two-agent environment, while model-free agents learn faster alone.

3.3 Learning timescales

How long do learning agents require to converge to their final strategies? As we expected, model-based learning dynamics converge faster than model-free learning. Figure 4 reveals that the difference in learning speed is about 1-2 orders of magnitude, as shown by the extent of the logarithmic y-axis in the plots. For most levels of environmental uncertainty, model-free learning in both single-agent and two-agent environments requires between 10^2 and 10^5 time steps, whereas model-based learning occurs between 10^1 and 10^3-10^4 time steps.

Interestingly, we observe critical transitions at the strategy transition points identified in the previous section. Here, learning slows dramatically: the convergence timescales peak, signaling a critical slowing down. This effect appears for both model-based and model-free dynamics and reflects agents being caught between two competing strategies. When the expected benefits of alternative actions are similar, updates become small, and convergence is prolonged. For example, a second-acting agent deciding between "copy-when-uncertain" and relying on its private signal will transition only slowly from choosing A at observation aB and B at bA to the opposite choices. Model-free learners show an additional peak at very low observational uncertainty σ because strong signals (A and B) occur so rarely that experience at those observations is sparse, which further increases indecision and slows convergence.

Generally, model-free agents learn faster alone, whereas model-based agents learn faster together. Only at the critical points do the model-based agents learn more slowly together than alone, as they have to undergo a critical transition to a different social learning strategy. Outside these critical points, two model-based agents learn about twice as fast as a single model-based agent, whereas a single model-free agent learns about twice as fast as two model-free agents – as, e.g., the parallel lines in Figure 4 A) for $\sigma > 1.5$ indicate. This model-free speed disadvantage arises from the stationary observation frequency weighting (P_X^{i,o^i}) in the model-free update rule: in the two-agent environment, each acting observation is visited less frequently in stationarity (because each agent is inactive during approximately half of all steps) so updates are smaller per step. Model-based agents are insulated from this effect because their updates use analytically computed expected rewards regardless of how often each observation is visited. The model-based speed advantage has a distinct origin. Under the assumption that social observation is temporally free, each social round

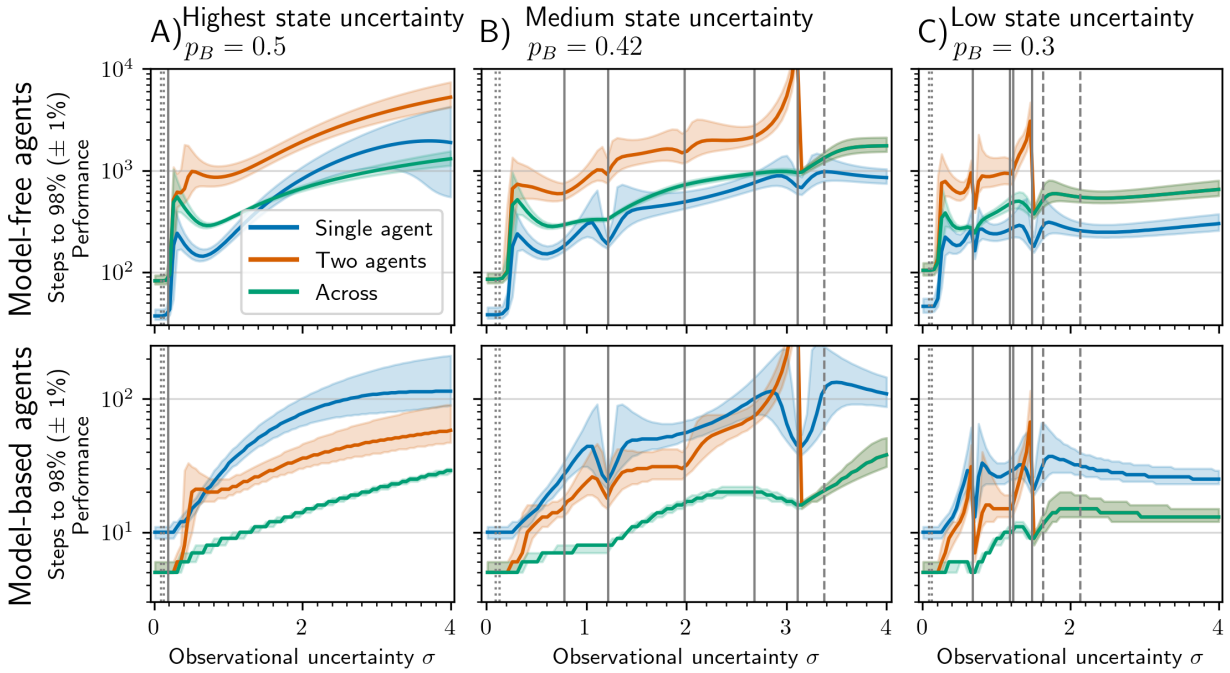

Figure 5: **Learning time steps** required until 98% (± 1%) of the final reward at strategy convergence is reached versus observational uncertainty σ for three different state uncertainties $p_B = 0.5$ (A), $p_B = 0.42$ (B), $p_B = 0.3$ (C). The upper panels show model-free learning dynamics, the lower panels show the model-based ones. All trajectories started from uniformly random strategies, as we saw in Figure 4 that multiple initial conditions do not have a significant impact on the results. Blue (vermillion) colors indicate the single-agent (two-agent) environment. The green color indicates the time steps required for the two learners to reach the respective performance level of the single agent. Most of the time spent on strategy convergence is devoted to achieving outperformance over independent learning; reaching near-optimal reward takes an order of magnitude fewer steps than full strategy convergence.

(in which both agents act sequentially) occupies the same wall-clock time as a single-agent step yet generates two reward-earning interactions. An agent with a world model can compute expected rewards over the full joint dynamics, directly exploiting the other agent's choice,and thus, achieving an approximately two-fold increase in effective learning rate. Without a world model, an agent updates only in proportion to its own individual observation encounters, which are not doubled by the social structure. The collective richness of the social round remains cognitively out of reach. Thus, when having a reliable internal model of the environment, learning together not only leads to better results, but it is also faster.

However, what if such an internal model is not available? How long does it take until a decent amount of the final reward is achieved? Figure 5 shows the time steps required to achieve 98% of the final reward (dark lines). The shaded regions show the times required to achieve between 99% (at the top) and 97% (at the bottom) of the final reward. Blue colors indicate single-agent learning, vermillion colors indicate two-agent learning, and green colors indicate the time steps required for two-agent learning to outperform 98% (± 1%) of the final reward achieved by single-agent learning.

Overall, it takes all learning dynamics about an order of magnitude less time to achieve most of the final reward outperformance than to fully converge in strategy (compare the y-scales of Figure 5 and Figure 4). Even some of the critical transitions in learning time scales identified in the strategy convergence analysis are smoothed out when considering the time steps required to achieve most of the final reward outperformance. Most interestingly, however, is the time required for two-agent learning to achieve most of the performance of single-agent learning (green lines in Figure 5). The model-based agents require only between 3 and 30 steps across all levels of environmental uncertainty. The model-free agents require longer, but in the order of magnitude comparable to how much the single model-free agents require to achieve most of their final reward. Thus, most of the time spent converging in strategy is devoted to achieving outperformance over independent learning.

Taken together, we find that learning together not only yields better results under environmental uncertainty but can also significantly speed up the learning process – provided that agents have a reliable internal model of the environment.

When such an internal model is not available, learning together still yields better results, but the learning speed advantage is lost. However, even model-free learning does not take significantly longer than independent learning to achieve most of the final reward that an independently learning agent achieves.

4 Discussion

In this work, we have systematically investigated how environmental uncertainty shapes the timescales of learning social learning strategies through pure reward-based learning. We find that learning together through reward-based learning not only yields results up to 14% better than independent learning across wide ranges of environmental uncertainty, regardless of whether agents possess internal world models. If agents possess reliable internal world models, learning together typically occurs also faster than learning alone, despite the increased complexity of the strategy the agents must learn. Without such internal world models, learning together is slower than learning alone. Yet, it takes about the same amount of time to achieve most of an independent agent's final performance. Thus, learning together poses no significant disadvantage in terms of learning speed. These findings highlight an often-overlooked dimension of the remarkable success of social learning strategies: they not only can yield better outcomes but also can accelerate the learning process itself. This finding helps explain why social learning is so widespread across diverse species and ecological contexts [Boyd et al., 2011].

One of the findings we were most surprised about is the rich diversity of social learning strategies that emerged in our comparably idealized model from pure reward-based learning. We designed our model with the intention of being the 'simplest' model that could show the emergence of a 'copy-when-uncertain' strategy. Indeed, this is what we find in the typical case of full environmental state uncertainty. However, when the environment is not fully uncertain, we find that more nuanced social learning strategies emerge, such as a 'copy-when-uncertain-about-the-less-likely-state' strategy and a 'follow-the-leader-in-the-more-likely-state' strategy. These strategies are not pre-programmed heuristics but emerge from our domain-general reward-learning mechanism, highlighting the flexibility and adaptability of learned social-learning strategies [Kendal et al., 2018]. Reinforcement learning was originally conceived as a single-agent learning process [Sutton and Barto, 2018]. Thus, there is an ongoing search for the best way to incorporate 'the social' into it, as evidenced by previous works devising various learning processes that incorporate sociality either via stimulus enhancement [Galef, 1988], value shaping [Najar et al., 2020], decision biasing [Toyokawa et al., 2019], social generalization [Witt et al., 2024], or social features [Schultner et al., 2025a]. Our learning dynamics most closely resemble the model by Schultner et al. [2025a]. Thus, our results suggest that no explicit incorporation of 'the social' into a reinforcement learning update is required beyond observing social features or signals. As we have shown, complex social learning strategies can emerge out of pure reward associations, resonating with recent calls in evolutionary game theory to move beyond hand-picked strategy sets toward learning-based strategy generation [Garcia and Traulsen, 2025].

The transitions between these learned social learning strategies also robustly reveal another interesting phenomenon: critical slowing down at the transition points. This is consistent with the general theory of critical transitions [Scheffer et al., 2009]: near a bifurcation point, the dominant eigenvalue of the system's Jacobian approaches zero, and convergence times diverge. We hypothesize that each strategy transition corresponds to such a bifurcation of the learning dynamics, and we leave a formal eigenvalue analysis for future work. At these points, learning timescales peak, indicating that agents take multiple orders of magnitude longer to converge to their final strategies. However, this time is spent achieving only the last few percent of the final rewards. Thus, from a practical perspective, these critical slowing down points may not pose significant hindrances to effective decisions, as agents can achieve most of their performance benefits relatively quickly. In contrast, being in a critical transition offers the advantage of flexible adaptation to changing environmental conditions, as the strategies are not yet fully stabilized [Beggs, 2007, Klamser and Romanczuk, 2021]. Importantly, this phenomenon is not an artifact of our deterministic dynamics. The bifurcation structure is a property of the underlying reward landscape, and stochastic RL algorithms would exhibit the same qualitative behavior near strategy transitions, manifesting as longer expected convergence times and increased variance in learning trajectories.

These findings provide insights for designing and understanding natural, artificial, and hybrid multi-agent systems, especially regarding the cognitive requirements for social learning. The fact that model-free agents – without an internal environmental model and only simple associative learning – can acquire sophisticated, multi-layered social learning strategies suggests that complex social behavior need not require complex cognition. Even organisms with relatively simple nervous systems might develop effective social learning strategies through basic reward associations [Leadbeater and Chittka, 2007, Chittka and Rossi, 2022]. This could explain why adaptive social learning appears in taxonomically diverse groups, from insects to mammals. However, our finding that model-based learners acquire social strategies much faster than model-free learners suggests a cognitive advantage for organisms that can build internal models of their environments. Mechanistically, this advantage is specific to the social setting. A world model enables an agent to leverage the collective dynamics of social interaction, pooling the doubled reward signal that each social round provides,

rather than being confined to learning from its own directly experienced observations at individual encounter rates. In rapidly changing environments or during critical developmental periods, the speed of learning could be decisive for survival. This provides a potential adaptive explanation for why more cognitively sophisticated organisms evolved the capacity for model-based learning: not just to achieve better final performance, but to adapt quickly enough to survive in dynamic environments [Suzuki and Arita, 2004, Fernando et al., 2018]. Regarding AI systems, this also means that their architecture has profound implications for collective learning speed. AI systems with explicit world models [Ha and Schmidhuber, 2018] should accelerate collective learning when they interact with humans or other AI agents. By contrast, purely reactive systems without internal models – analogous to our model-free learners – may slow down collective adaptation even as they eventually reach good solutions. This finding is supported by empirical research in multi-agent reinforcement learning. Purely model-free agents have been shown to fail to adopt social learning strategies without an auxiliary model-based learning component, whereas agents with world models develop generalised social learning policies and transfer them to novel environments [Ndousse et al., 2021]. This suggests that investments in model-based AI architectures may pay off not just in individual AI performance but also in AI agents' capacity for cultural accumulation [Cook et al., 2024] and emergent social behaviour [Köster et al., 2025]. More broadly, our findings contribute a micro-founded learning mechanism to emerging accounts of machine culture and machine behaviour in hybrid human-AI populations [Rahwan et al., 2019, Brinkmann et al., 2023], showing how agents embedded in shared environments develop and transmit social learning strategies through pure reward associations, without any explicit programming of cultural norms.

Methodologically, our work demonstrates that evolutionary and learning-dynamics approaches based on replicator-like dynamics [Hofbauer and Sigmund, 1998, Fudenberg and Levine, 2016, Barfuss, 2020a] can investigate the emergence of sociality from pure reward-based learning. Specifically, our environment model can serve as a milestone from where further questions can be investigated, such as different preference structures, asymmetric information, noisy or delayed social observations, larger group sizes including varied network topologies, and non-stationary changes in more complex environments. To our knowledge, this work is the first to apply the concepts of model-free and model-based learning to replicator-like learning dynamics [Barfuss, 2020b]. Classic reinforcement learning dynamics were using the model-based variant [Tuyls et al., 2003, Sato and Crutchfield, 2003, Barfuss et al., 2019][Leslie and Collins, 2005], without being aware of that interpretation [Kaisers and Tuyls, 2011]. More recent works incorporate choice frequencies into the reinforcement signal, as our model-free version [Hu et al., 2022, Goll et al., 2024], but have not made the connection to model-free learning explicit. Thus, our work contributes to the bridges between the fields of multi-agent reinforcement learning and evolutionary game theory, opening new avenues for cross-disciplinary research[Madhushani Sehwag et al., 2025, Barfuss et al., 2025]. Recent work in evolutionary game theory has investigated how rational decision-making strategies themselves evolve rather than being assumed a priori [Salahshour, 2025, Salahshour and Couzin, 2025], offering a complementary perspective to our reward-learning account of emergent strategies. Future work should endogenize the development of internal world models through experience rather than assuming perfect internal models from the start. While these deterministic dynamics provide analytical tractability, they abstract away potentially important stochastic effects [Galla, 2009, Barfuss and Meylahn, 2023, Rudd-Jones et al., 2025]. Further research is required to explore how stochasticity in learning processes influences the emergence of social learning strategies. We expect the qualitative speed ordering model-based agents learning faster together, model-free agents learning faster alone to be robust to stochastic sampling, since the underlying mechanism is informational rather than algorithmic: any model-based algorithm exploiting the joint dynamics will benefit from the doubled reward signal per social round, while any model-free algorithm updating in proportion to empirical visit rates will be penalized by the sparser individual observation frequencies in the two-agent environment. We also consider it conceivable that some of our findings, such as the existence of a critical slowing down in learning, could serve as hypotheses to be tested experimentally with human subjects in controlled laboratory settings.

In conclusion, our work demonstrates that sophisticated social learning strategies need not be inherited, rational, or preprogrammed — they can arise quickly whenever agents seek and learn from reward in uncertain environments.

Acknowledgments. WB acknowledges financial support from the Cooperative AI Foundation. RPM was supported by a UKRI Future Leaders Fellowship (MR/X036863/1) and Templeton World Charity Foundation Inc. (TWCF202120647).

Code availability. All computer code was written in Python and is contained in reproducible Supplementary Information. We will make it openly available upon acceptance of this manuscript.

References

Alberto Acerbi, Alex Mesoudi, and Marco Smolla. *Individual-Based Models of Cultural Evolution*. 2022.

Kenichi Aoki and Marcus W. Feldman. Evolution of learning strategies in temporally and spatially variable environments: A review of theory. *Theoretical Population Biology*, 91:3–19, February 2014. ISSN 0040-5809. doi: 10.1016/j.tpb.2013.10.004.

Sara Arganda, Alfonso Pérez-Escudero, and Gonzalo G. de Polavieja. A common rule for decision making in animal collectives across species. *Proceedings of the National Academy of Sciences*, 109(50):20508–20513, December 2012. doi: 10.1073/pnas.1210664109.

Abhijit V. Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817, 1992. doi: 10.2307/2118364.

Wolfram Barfuss. Towards a unified treatment of the dynamics of collective learning. In *Challenges and Opportunities for Multi-Agent Reinforcement Learning, AAAI Spring Symposium*, 2020a.

Wolfram Barfuss. Reinforcement Learning Dynamics in the Infinite Memory Limit. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 1768–1770, Richland, SC, May 2020b. International Foundation for Autonomous Agents and Multiagent Systems.

Wolfram Barfuss. Dynamical systems as a level of cognitive analysis of multi-agent learning. *Neural Computing and Applications*, 34(3):1653–1671, February 2022. ISSN 1433-3058. doi: 10.1007/s00521-021-06117-0.

Wolfram Barfuss and Richard P. Mann. Modeling the effects of environmental and perceptual uncertainty using deterministic reinforcement learning dynamics with partial observability. *Physical Review E*, 105(3):034409, March 2022. doi: 10.1103/PhysRevE.105.034409.

Wolfram Barfuss and Janusz M. Meylahn. Intrinsic fluctuations of reinforcement learning promote cooperation. *Scientific Reports*, 13(1):1309, January 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-27672-7.

Wolfram Barfuss, Jonathan F. Donges, and Jürgen Kurths. Deterministic limit of temporal difference reinforcement learning for stochastic games. *Physical Review E*, 99(4):043305, April 2019. doi: 10.1103/PhysRevE.99.043305.

Wolfram Barfuss, Jessica Flack, Chaitanya S. Gokhale, Lewis Hammond, Christian Hilbe, Edward Hughes, Joel Z. Leibo, Tom Lenaerts, Naomi Leonard, Simon Levin, Udari Madhushani Sehwag, Alex McAvoy, Janusz M. Meylahn, and Fernando P. Santos. Collective cooperative intelligence. *Proceedings of the National Academy of Sciences*, 122 (25):e2319948121, June 2025. doi: 10.1073/pnas.2319948121.

John M Beggs. The criticality hypothesis: How local cortical networks might optimize information processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1864): 329–343, August 2007. ISSN 1364-503X. doi: 10.1098/rsta.2007.2092.

Timothy E. J. Behrens, Laurence T. Hunt, Mark W. Woolrich, and Matthew F. S. Rushworth. Associative learning of social value. *Nature*, 456(7219):245–249, November 2008. ISSN 1476-4687. doi: 10.1038/nature07538.

Clémence Bergerot, Wolfram Barfuss, and Pawel Romanczuk. Moderate confirmation bias enhances decision-making in groups of reinforcement-learning agents. *PLOS Computational Biology*, 20(9):e1012404, September 2024. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1012404.

Martino Bernasconi, Federico Cacciamani, Simone Fioravanti, Nicola Gatti, and Francesco Trovò. The evolutionary dynamics of soft-max policy gradient in multi-agent settings. *Theoretical Computer Science*, 1027:115011, 2025. doi: 10.1016/j.tcs.2024.115011.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, 100(5):992–1026, October 1992. ISSN 0022-3808, 1537-534X. doi: 10.1086/261849.

Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary Dynamics of Multi-Agent Learning: A Survey. *Journal of Artificial Intelligence Research*, 53:659–697, August 2015. ISSN 1076-9757. doi: 10.1613/jair.4818.

Tilman Börgers and Rajiv Sarin. Learning Through Reinforcement and Replicator Dynamics. *Journal of Economic Theory*, 77(1):1–14, November 1997. ISSN 00220531. doi: 10.1006/jeth.1997.2319.

Diana Borsa, Nicolas Heess, Bilal Piot, Siqi Liu, Leonard Hasenclever, Remi Munos, and Olivier Pietquin. Observational Learning by Reinforcement Learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pages 1117–1124, Richland, SC, May 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-6309-9.

Robert Boyd and Peter J. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, June 1988. ISBN 978-0-226-06933-3.

Robert Boyd, Peter J. Richerson, and Joseph Henrich. The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, 108(supplement_2):10918–10925, June 2011. doi: 10.1073/pnas.1100290108.

Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F. Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L. Griffiths, Joseph Henrich, Joel Z. Leibo, Richard McElreath, Pierre-Yves Oudeyer, Jonathan Stray, and Iyad Rahwan. Machine culture. *Nature Human Behaviour*, 7(11): 1855–1868, November 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01742-2.

Lars Chittka and Natacha Rossi. Social cognition in insects. *Trends in Cognitive Sciences*, 26(7):578–592, July 2022. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2022.04.001.

Jonathan Cook, Chris Lu, Edward Hughes, Joel Z. Leibo, and Jakob Foerster. Artificial Generational Intelligence: Cultural Accumulation in Reinforcement Learning. In *Advances in Neural Information Processing Systems 37*, volume 37, pages 59689–59715, December 2024. doi: 10.52202/079017-1907.

Lukas Danwitz and Bettina von Helversen. Observational learning of exploration-exploitation strategies in bandit tasks. *Cognition*, 259:106124, 2025. doi: 10.1016/j.cognition.2025.106124.

Dominik Deffner, Vivien Kleinow, and Richard McElreath. Dynamic social learning in temporally and spatially variable environments. *Royal Society Open Science*, 7(12):200734, December 2020. ISSN 2054-5703. doi: 10.1098/rsos.200734.

Dominik Deffner, Natalia Fedorova, Jeffrey Andrews, and Richard McElreath. Bridging theory and data: A computational workflow for cultural evolution. *Proceedings of the National Academy of Sciences*, 121(48):e2322887121, November 2024. doi: 10.1073/pnas.2322887121.

Charles Efferson, Rafael Lalive, Peter J. Richerson, Richard McElreath, and Mark Lubell. Conformists and mavericks: The empirics of frequency-dependent cultural transmission. *Evolution and Human Behavior*, 29(1):56–64, January 2008. ISSN 1090-5138. doi: 10.1016/j.evolhumbehav.2007.08.003.

Chrisantha Fernando, Jakub Sygnowski, Simon Osindero, Jane Wang, Tom Schaul, Denis Teplyashin, Pablo Sprechmann, Alexander Pritzel, and Andrei Rusu. Meta-learning by the Baldwin effect. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '18, pages 1313–1320, New York, NY, USA, July 2018. Association for Computing Machinery. doi: 10.1145/3205651.3208249.

Drew Fudenberg and David K. Levine. Whither Game Theory? Towards a Theory of Learning in Games. *Journal of Economic Perspectives*, 30(4):151–170, November 2016. ISSN 0895-3309. doi: 10.1257/jep.30.4.151.

Bennett G. Galef. Imitation in Animals: History, Definition, and Interpretation of Data from the Psychological Laboratory. In *Social Learning*. Psychology Press, 1988.

Tobias Galla. Intrinsic Noise in Game Dynamical Learning. *Physical Review Letters*, 103(19):198702, November 2009. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.103.198702.

Julian Garcia and Arne Traulsen. Picking Strategies in Games of Cooperation. *Proceedings of the National Academy of Sciences*, 122(25):e2319925121, June 2025. doi: 10.1073/pnas.2319925121.

David Goll, Jobst Heitzig, and Wolfram Barfuss. Deterministic Model of Incremental Multi-Agent Boltzmann Q-Learning: Transient Cooperation, Metastability, and Oscillations, December 2024.

David Ha and Jürgen Schmidhuber. World Models. *World Models*, 1(1):e10, March 2018. doi: 10.5281/zenodo.12076 31.

Seungwoong Ha and Hawoong Jeong. Social learning spontaneously emerges by searching optimal heuristics with deep reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 12319–12338. PMLR, July 2023.

Cecilia Heyes. What's social about social learning? *Journal of Comparative Psychology*, 126(2):193–202, 2012. ISSN 1939-2087. doi: 10.1037/a0025180.

Cecilia Heyes. Blackboxing: Social learning strategies and cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693):20150369, May 2016. ISSN 0962-8436. doi: 10.1098/rstb.2015.0369.

Cecilia Heyes and John M. Pearce. Not-so-social learning strategies. *Proceedings of the Royal Society B: Biological Sciences*, 282(1802):20141709, March 2015. ISSN 0962-8452. doi: 10.1098/rspb.2014.1709.

Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1 edition, May 1998. ISBN 978-0-521-62365-0 978-0-521-62570-8 978-1-139-17317-9. doi: 10.1017/CBO9781139173179.

Shuyue Hu, Chin-Wing Leung, Ho-fung Leung, and Harold Soh. The Dynamics of Q-learning in Population Games: A Physics-inspired Continuity Equation Model. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pages 615–623, Richland, SC, May 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-9213-6.

Michael Kaisers and Karl Tuyls. FAQ-learning in matrix games: Demonstrating convergence near Nash equilibria, and bifurcation of attractors in the battle of sexes. In *Proceedings of the 13th AAAI Conference on Interactive Decision Theory and Game Theory*, AAAIWS'11-13, pages 36–42. AAAI Press, January 2011.

Rachel L. Kendal, Neeltje J. Boogert, Luke Rendell, Kevin N. Laland, Mike Webster, and Patricia L. Jones. Social learning strategies: Bridge-building between fields. *Trends in cognitive sciences*, 22(7):651–665, 2018.

Pascal P. Klamser and Pawel Romanczuk. Collective predator evasion: Putting the criticality hypothesis to the test. *PLOS Computational Biology*, 17(3):e1008832, March 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008832.

Raphael Köster, Edgar A. Duéñez-Guzmán, William A. Cunningham, and Joel Z. Leibo. Tabula Rasa Agents Display Emergent In-group Behavior. *Proceedings of the National Academy of Sciences*, 122(25):e2319947121, June 2025. doi: 10.1073/pnas.2319947121.

Kevin N. Laland. Social learning strategies. *Learning & Behavior*, 32(1):4–14, February 2004. ISSN 1543-4508. doi: 10.3758/BF03196002.

Ellouise Leadbeater and Lars Chittka. Social Learning in Insects — From Miniature Brains to Consensus Building. *Current Biology*, 17(16):R703–R713, August 2007. ISSN 0960-9822. doi: 10.1016/j.cub.2007.06.012.

David S. Leslie and Edmund J. Collins. Individual q-learning in normal form games. *SIAM Journal on Control and Optimization*, 44(2):495–514, 2005. doi: 10.1137/S0363012903437976.

Udari Madhushani Sehwag, Alex McAvoy, and Joshua B. Plotkin. Collective artificial intelligence and evolutionary dynamics. *Proceedings of the National Academy of Sciences*, 122(25):e2505860122, June 2025. doi: 10.1073/pnas.2505860122.

Richard P. Mann. Collective decision making by rational individuals. *Proceedings of the National Academy of Sciences*, 115(44):E10387–E10396, October 2018. doi: 10.1073/pnas.1811964115.

Richard P. Mann. Collective decision-making by rational agents with differing preferences. *Proceedings of the National Academy of Sciences*, 117(19):10388–10396, May 2020. doi: 10.1073/pnas.2000840117.

T. J. H. Morgan, L. E. Rendell, M. Ehn, W. Hoppitt, and K. N. Laland. The evolutionary basis of human social learning. *Proceedings of the Royal Society B: Biological Sciences*, 279(1729):653–662, July 2011. ISSN 0962-8452. doi: 10.1098/rspb.2011.1172.

Anis Najar, Emmanuelle Bonnet, Bahador Bahrami, and Stefano Palminteri. The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLOS Biology*, 18(12):e3001028, December 2020. ISSN 1545-7885. doi: 10.1371/journal.pbio.3001028.

Kamal K. Ndousse, Douglas Eck, Sergey Levine, and Natasha Jaques. Emergent Social Learning via Multi-agent Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7991–8004. PMLR, July 2021.

Andreas Olsson, Ewelina Knapska, and Björn Lindström. The neural and computational systems of social learning. *Nature Reviews Neuroscience*, 21(4):197–212, April 2020. ISSN 1471-0048. doi: 10.1038/s41583-020-0276-4.

Alfonso Pérez-Escudero and Gonzalo G. de Polavieja. Collective Animal Behavior from Bayesian Estimation and Probability Matching. *PLOS Computational Biology*, 7(11):e1002282, November 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002282.

Charles Perreault, Cristina Moya, and Robert Boyd. A Bayesian approach to the evolution of social learning. *Evolution and Human Behavior*, 33(5):449–459, September 2012. ISSN 1090-5138. doi: 10.1016/j.evolhumbehav.2011.12.007.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M.

Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. Machine behaviour. *Nature*, 568(7753):477–486, April 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1138-y.

L. Rendell, R. Boyd, D. Cownden, M. Enquist, K. Eriksson, M. W. Feldman, L. Fogarty, S. Ghirlanda, T. Lillicrap, and K. N. Laland. Why Copy Others? Insights from the Social Learning Strategies Tournament. *Science*, 328(5975): 208–213, April 2010. doi: 10.1126/science.1184719.

Alan R. Rogers. Does biology constrain culture? *American Anthropologist*, 90(4):819–831, 1988. doi: 10.1525/aa.1988.90.4.02a00030.

James Rudd-Jones, María Pérez-Ortiz, and Mirco Musolesi. Understanding Individual Decision-Making in Multi-Agent Reinforcement Learning: A Dynamical Systems Approach, December 2025.

Mohammad Salahshour. Perceptual rationality: An evolutionary game theory of perceptually rational decision-making. *Royal Society Open Science*, 12(10), 2025. doi: 10.1098/rsos.251125.

Mohammad Salahshour and Iain D. Couzin. Evolution of altruistic rationality provides a solution to social dilemmas via rational reciprocity. *Physical Review Research*, 7(3):033211, 2025. doi: 10.1103/sz5b-j75y.

Yuzuru Sato and James P. Crutchfield. Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E*, 67(1):015206, January 2003. ISSN 1063-651X, 1095-3787. doi: 10.1103/PhysRevE.67.015206.

Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. Van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009. doi: 10.1038/nature08227.

David Schultner, Lucas Molleman, and Björn Lindström. Feature-based reward learning shapes human social learning strategies. *Nature Human Behaviour*, pages 1–16, July 2025a. ISSN 2397-3374. doi: 10.1038/s41562-025-02269-4.

David Schultner, Lucas Molleman, and Björn Lindström. Reward is enough for social learning. *Trends in Cognitive Sciences*, 29(9):787–789, September 2025b. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2025.06.012.

Anna Sigalou and Richard P Mann. Evolutionary stability of social interaction rules in collective decision-making. *Physical Biology*, 20(4):045003, May 2023. ISSN 1478-3975. doi: 10.1088/1478-3975/acd297.

Hidezo Suganuma, Kentaro Katahira, Hisashi Ohtsuki, and Tatsuya Kameda. How social learning enhances—or undermines—efficiency and flexibility in collective decision-making under uncertainty. *Proceedings of the National Academy of Sciences*, 122(48):e2516827122, December 2025. doi: 10.1073/pnas.2516827122.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03924-6.

Reiji Suzuki and Takaya Arita. Interactions between learning and evolution: The outstanding strategy generated by the Baldwin effect. *Bio Systems*, 77(1-3):57–71, November 2004. ISSN 0303-2647. doi: 10.1016/j.biosystems.2004.04.002.

Max Taylor-Davies, Neil Bramley, and Christopher G. Lucas. A Rational Framework for Group-Based Selective Social Learning. *Open Mind*, 9:677–708, May 2025. ISSN 2470-2986. doi: 10.1162/opmi_a_00205.

Ulf Toelch, Matthew J. Bruce, Lesley Newson, Peter J. Richerson, and Simon M. Reader. Individual consistency and flexibility in human social information use. *Proceedings of the Royal Society B: Biological Sciences*, 281(1776): 20132864, February 2014. ISSN 0962-8452. doi: 10.1098/rspb.2013.2864.

Wataru Toyokawa and Wolfgang Gaissmaier. Conformist social learning leads to self-organised prevention against adverse bias in risky decision making. *eLife*, 11:e75308, May 2022. ISSN 2050-084X. doi: 10.7554/eLife.75308.

Wataru Toyokawa, Andrew Whalen, and Kevin N. Laland. Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, 3(2):183–193, February 2019. ISSN 2397-3374. doi: 10.1038/s41562-018-0518-x.

Alan N. Tump, Dominik Deffner, Timothy J. Pleskac, Pawel Romanczuk, and Ralf H. J. M. Kurvers. A Cognitive Computational Approach to Social and Collective Decision-Making. *Perspectives on Psychological Science*, 19(2): 538–551, March 2024. ISSN 1745-6916. doi: 10.1177/17456916231186964.

Matthew A. Turner, Cristina Moya, Paul E. Smaldino, and James Holland Jones. The form of uncertainty affects selection for social learning. *Evolutionary Human Sciences*, 5:e20, January 2023. ISSN 2513-843X. doi: 10.1017/ehs.2023.11.

Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. A selection-mutation model for q-learning in multi-agent systems. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems - AAMAS '03*, page 693, Melbourne, Australia, 2003. ACM Press. ISBN 978-1-58113-683-8. doi: 10.1145/860575.860687.

Alexandra Witt, Wataru Toyokawa, Kevin N. Lala, Wolfgang Gaissmaier, and Charley M. Wu. Humans flexibly integrate social information despite interindividual differences in reward. *Proceedings of the National Academy of Sciences*, 121(39):e2404928121, September 2024. doi: 10.1073/pnas.2404928121.

Charley M. Wu, Dominik Deffner, Björn Kahl, Björn Meder, Mark K. Ho, and Ralf H. J. M. Kurvers. Adaptive mechanisms of social and asocial learning in immersive collective foraging. *Nature Communications*, 16(1):3539, 2025. doi: 10.1038/s41467-025-58365-6.

Lei Zhang and Jan Gläscher. A brain network supporting social influences in human decision-making. *Science Advances*, 6(34):eabb4159, August 2020. doi: 10.1126/sciadv.abb4159.