Meta-control of social learning strategies

Anil Yaman, Nicolas Bredeche, Onur Ç Aylak, Joel Z Leibo, Sang Wan Lee

Meta-control of social learning strategies

Anil Yaman[1]*, Nicolas Bredeche[2], Onur Çaylak[3], Joel Z Leibo[4], Sang Wan Lee[1]

[1]Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
[2]Sorbonne Université, Paris, France
[3]Eindhoven University of Technology, Eindhoven, the Netherlands
[4]DeepMind, London, UK

Abstract

Social learning, copying other's behavior without actual experience, offers a cost-effective means of knowledge acquisition. However, it raises the fundamental question of which individuals have reliable information: successful individuals versus the majority. The former and the latter are known respectively as success-based and conformist social learning strategies. We show here that while the success-based strategy fully exploits the benign environment of low uncertainly, it fails in uncertain environments. On the other hand, the conformist strategy can effectively mitigate this adverse effect. Based on these findings, we hypothesized that meta-control of individual and social learning strategies provides effective and sample-efficient learning in volatile and uncertain environments. Simulations on a set of environments with various levels of volatility and uncertainty confirmed our hypothesis. The results imply that meta-control of social learning affords agents the leverage to resolve environmental uncertainty with minimal exploration cost, by exploiting others' learning as an external knowledge base.

1 Introduction

Learning is one of the basic requirements for animal survival. The ultimate goal of learning is to acquire reliable knowledge from a limited amount of interactions with the environment. However, the environment is often uncertain and volatile, making it difficult to learn.

Decades of studies found that animals have multiple learning strategies. For example, an animal can learn associations between environmental cues and outcomes (Pavlovian learning) or learn action-outcome associations (model-free reinforcement learning). It is a simple strategy but less adaptive to environmental changes because the action-outcome associations are gradually updated based on experience. A more sophisticated strategy is to learn the internal model of the environmental

*Corresponding author: anilyaman@kaist.ac.kr

structure and to use this information to quickly perform actions in a more predictive manner (model-based reinforcement learning). While this strategy can accelerate the adaptation, encoding the internal model of the environment requires additional memory and computations [37, 46].

Recent studies in neuroscience suggest that the brain can find a compromise between these learning strategies via a process called meta-control [9, 61, 47, 17]. Meta-control is based on the premise that learning strategies have different levels of sensitivity to environment variability and this can be measured by perceptual uncertainty concerning the association of actions and rewards [9]. Thus, perceptual uncertainty can be used to arbitrate between learning strategies. For example, in a stable environment, the brain prefers to use a sample-efficient model-based strategy, followed by a gradual transition to a computationally-efficient model-free learning strategy [20]. On the other hand, in environments with high perceived uncertainty, model-free learning are preferred over model-based learning because they are less susceptible to environmental uncertainty [39]. Accumulating evidence suggests that a part of the prefrontal cortex implements meta-control of various learning strategies, which provides a cost-effective solution to environmental uncertainty [37, 32, 49]. Ultimately, computational models of the brain's meta-control principle should find a way to efficiently avoid complications arising from environmental variability [36].

Taking full advantage of the brain's meta-control capability of learning strategies, this paper proposes a meta-control approach to social learning, which we term *meta-social learning*. The proposed method aims to resolve environmental uncertainty by arbitrating between individual learning and two different social learning strategies, each of which exhibits a different uncertainty-sensitive performance-cost profile.

Both the individual and social learning strategies play a crucial role in learning as a population. Innovations are usually made by individual learning (IL) and spread throughout the population via social learning (SL) [23, 10, 19]. However, these strategies involve advantages and drawbacks, suggesting the need to trade social learning strategies off with individual learning [6, 24].

Individual learning can explore and discover useful innovations, however, it can be costly due to exploration, risk of injury, mortality, etc. [30]. It is only worthwhile to bear these costs if learning is required to adapt to the environmental changes. In static environments on the other hand, it would be unnecessary to pay these costs. To resolve this dilemma, one can suggest adapting the exploration rate so that it depends on the environment's variability (e.g. [59]). However, in this case, individual learners would still need to explore the action space by themselves to find the optimum behavior. This is inefficient if the optimum behavior was already discovered by other individuals in the population, and it can be readily copied. In that case it would be beneficial to make use of the knowledge explored by others in order to avoid paying the cost of exploring oneself.

In group-living animals in nature, social learning has evolved to take advantage of the exploration performed by others via copying their behavior, thereby reducing the cost of learning. Therefore, it does not involve these costs related to individual learning [25, 31, 60, 4]. On the other hand, social information can be less accurate since it depends on the observation of previously performed

behaviors (i.e. may be outdated in case of environment change). Moreover, it requires identifying the individuals with reliable knowledge to copy. The term social learning strategy (SLS) refers to any of a variety of methods by which individuals can choose others to copy [35, 31, 64, 25, 41].

The efficiency of individual and social learning strategies in stable and dynamic environments has been demonstrated through theoretical and empirical studies [22, 27, 1, 30, 29]. For instance in a computer tournament, Rendell *et al.* [50] noted the success of strategies that rely heavily on social learning over individual learning. They tested competing strategies on a fundamental decision-making problem known as the multi-armed bandit problem [58] (or $k-$armed bandit). Furthermore, they modeled a changing adaptive environment by adjusting the reward association of the arms during the task (it was a "restless" or "non-stationary" multi-armed bandit problem) [50, 53, 34, 21]. Despite these efforts, a fundamental issue in social learning still remains unaddressed: whether to use success-based social learning and copy the behavior of the successful individuals or simply follow the conformist strategy and copy the behavior that is the majority in the population. This poses a fundamental challenge for the individual, forcing them to confront a trade-off between performance and exploration cost [36].

To fully examine this issue, we performed an evolutionary analysis on individual learning and two social learning strategies, success-based and conformist [13, 31, 22, 30, 42]. Despite several investigations into these strategies, the effect of environment uncertainty on their performance has remained largely unaddressed, making it hard to connect to the meta-control idea in neuroscience. To fill this gap, we designed non-stationary multi-armed bandit tasks with variable amounts of uncertainty, for which we systematically manipulated the reward distributions. This setup has been used as an abstracted model of fundamental decision making processes in nature, such as foraging, predator avoidance, symbiosis, and mutualism [41, 8, 63], as well as human decision making processes, such as human-robot interactions, investment decisions in stock markets, consumer decision making, dynamics of social networks, etc.

Our simulations confirmed the view that while the success-based strategy is vulnerable to environmental uncertainty, the conformist strategy serves as an alternative that can effectively resolve the adverse effect of uncertainty. These results show that neither individual learning, success-based, nor conformist social learning strategies are on their own sufficient to achieve a optimal policy for lifetime learning. This view motivates us to hypothesize that there exists an ideal combination of these strategies to cope with the environment volatility and uncertainty. In doing so, we propose a meta-social learning strategy that uses estimated uncertainty to arbitrate between individual and social learning strategies.

To test our hypothesis, this model was pitted against a large set of algorithms implementing various strategies based on reinforcement learning [58], genetic algorithms [16] and neuroevolution [67, 56] in environments with various levels of volatility and uncertainty. The results demonstrate that the proposed model serves to achieve near-optimal lifetime learning in the sense that it resolves the performance-exploration cost dilemma. We also show that the version of the models

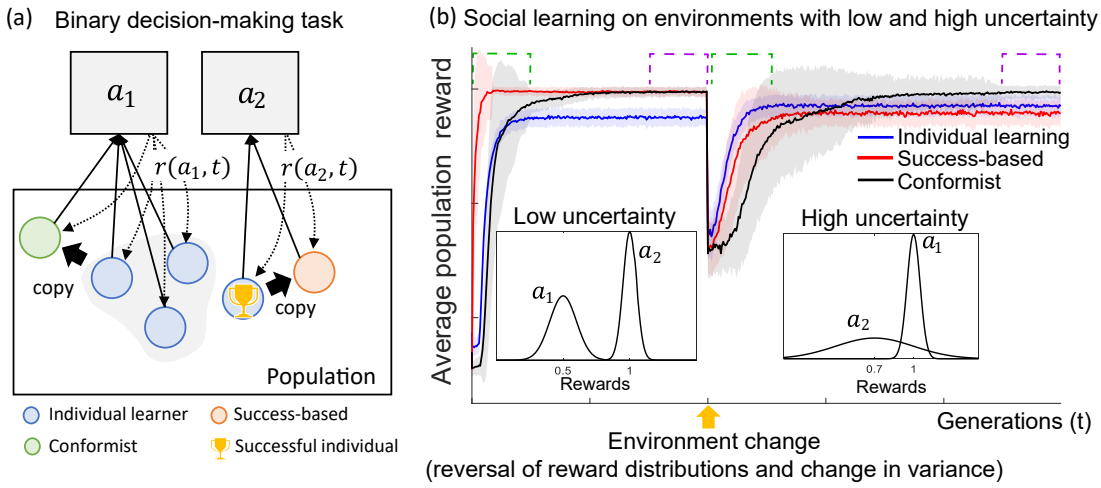


Figure 1: (a) A population of individuals perform a binary decision-making task based on individual and social learning strategies and collect their rewards ($r(a_j, t)$) based on their actions. The individual learners can perform their actions based on their decision models that can be improved by experience. The social learners use success-based or conformist strategies to copy the actions of successful individuals or the majority respectively. (b) Binary decision-making task (2-armed bandit) is iteratively performed for a certain period of time with specified reward distributions that are unknown to the individuals. At some point, an environment change occurs by changing the reward distributions (a.k.a reward reversal). In earlier stages of the process (initial and after environment change, shown in green dash lines), populations with success-based social learning strategy achieves higher average population reward faster relative to the conformist strategy. In later stages of the process (shown in purple dashed lines), populations with success-based social learning strategy achieves higher average population reward in environments with low uncertainty, whereas, populations with conformist social learners achieves higher average population reward in environments with high uncertainty.

implementing our hypothesis tends to have a higher ratios in the populations and survive longer relative to the others throughout our evolutionary analysis.

2 Results

2.1 Uncertainty-invariance in social learning

To examine whether and to what extent environmental uncertainty influences performance of individual and social learning strategies, we compared adaptation performance of various types of learning strategies in different levels of environmental uncertainty. We considered independent populations consisting of individual learners using the success-based social learning strategy, and individual learners with the conformist strategy. To model the evolutionary dynamics of these populations, we used two distinct approaches: (1) a mathematical model based on the replicator-mutator equa-

tion [33, 45], and (2) an agent-based evolutionary algorithm [16]. These approaches allow us to tract the change in the ratios of the individual and social learners within a certain environment.

In the former case, the change in the frequencies of individual/social learners selecting a given arm are modeled by a system of coupled first order differential equations. This approach has largely been used in evolutionary game theory [55, 26]. The fitness of the types of individuals was defined based on the rewards received. Note that the individual learners carry a constant computational cost of learning. To the contrary, the social learners avoid this issue by simply copying the others' choice, although they are deemed to endure a latency issue arising from the necessity of observing the past choices of the others.

In the latter case, we simulate the evolutionary process involving a population of individual and social learners, each of which were modelled separately. The individual learners have the capacity to improve their behavioral policy over time based on their experience. They were implemented using the ϵ-greedy algorithm [58] in which an exploration parameter ϵ is the probability per timestep of taking a random (uniformly distributed) action instead of taking the current greedy action with the highest average reward. Note that this exploration carries an extra cost when the individual is already making an optimal choice (see Section 4). The social learners, on the other hand, perform their actions by copying others with a certain level of latency. This process is repeated in each generation where all the individuals made their choices and receive their rewards. At the end of each generation cycle, individuals were sampled with replacement for the next generation proportional to their fitness.

An example illustration of a binary decision-making task given in Figure 1 (a) and (b). The reward distributions are parameterized by Gaussian distributions with a mean (μ) and standard deviation (σ). Since the performance of success-based social learning is contingent on correctly identifying agents making optimal choices, we hypothesized that the high uncertainty in the environment would make it hard to identify successful individuals, leading to the degradation of its learning outcomes. To test this, we designed novel tasks with different levels of uncertainty, controlled by the degree of overlap between reward distributions of different arms. The fitness values of the individuals are defined as the amount of rewards received following their choice. To assess the individuals' adaptation ability to environmental changes, at the midpoint of each simulation we reversed the association between arms and their reward distributions.

Figure 2 shows the performance comparison of the populations consisting of only individual learners, success-based and individual learners, and conformist and individual learners after the environment change (the average of the period between $t = [200, 250]$) and at the end of the process (average of the period between $t = [350, 400]$). To formally quantify the effect of uncertainty on performance, we also used critical difference (CD) diagrams[1]. The population dynamics throughout the evolutionary processes are shown in Figures 8 and 9 in Supplementary Material A.1. When

[1]The critical difference (CD) diagrams allow comparison of results from multiple strategies. They show the average ranks of the algorithms from the best to the worst. The algorithms that do not have significant rank difference are linked (post-hoc Nemenyi test [43] at $\alpha = 0.05$ [12]).

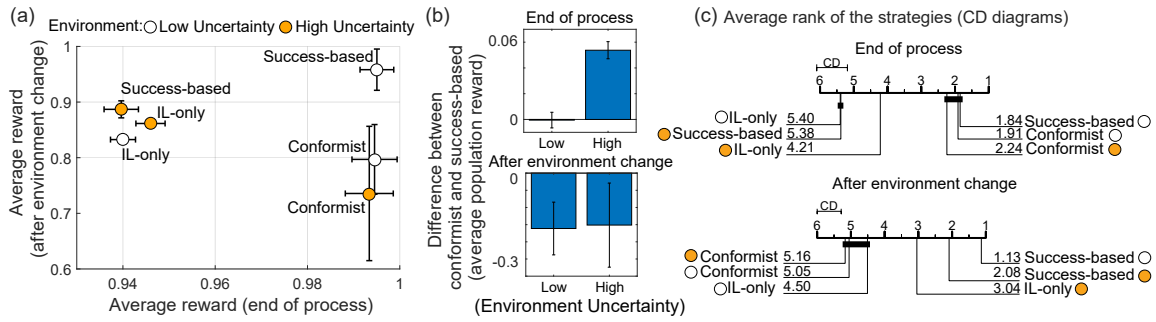


Figure 2: The success-based strategy shows the best performance in environments with low uncertainty, however, it suffers when there is uncertainty in the environment. On the other hand, the conformist strategy achieves similar performance independent of the environment uncertainty.

the uncertainty in the environment is low, the success-based social learning strategy shows the best performance in terms of adaptation after an environment change ($p < 0.01$; Wilcoxon rank-sum test [65]), and at the end of the process, the success-based and conformist strategies show similar performance, and they are superior to individual learning only. However, when the uncertainty in the environment is increased, populations with conformist social learners achieve higher average population reward ($p < 0.01$).

2.2 Uncertainty as a predictor of social learning performance

To further investigate the relationship between environmental uncertainty and social learning strategies, we measured the performance difference between the success-based and conformist strategy as a function of the amount of uncertainty in reward distributions. We refer to this uncertainty as the optimum distribution prediction uncertainty (ODPU) because it undermines the ability of the success-based strategy to correctly identify individuals making optimal choices. We computed the ODPU directly by the probability of receiving better rewards from the sub-optimal reward distributions. For example, if the ODPU is high, it is more likely to mistake an individual making sub-optimal choice as a successful individual and copy its action.

Considering a population consists of M and N individuals making choices to collect rewards from the environment associated with certain reward distributions. In this case, the ODPU depends not only on the sufficient statistics of the reward distributions but also on the size of the subgroup of individuals making optimal and sub-optimal choices, denoted by M and N, respectively. In Figure 3, we illustrate the performance difference between the success-based and conformist strategies as a function of the ODPU on two Gaussian reward distributions. The generalized version of the ODPU that can be applied to more than two reward distributions is provided in Section 4.4.

In Figures 3c and 3d, we show the ODPU when $M = 5$, $N = 95$ and $M = 50$, $N = 50$ depending on σ_1, σ_2 respectively. Overall, the smaller the number of individuals making optimal choices, the larger the ODPU. In addition, an increase in σ_1 and a decrease in σ_2 causes an increase in the ODPU.

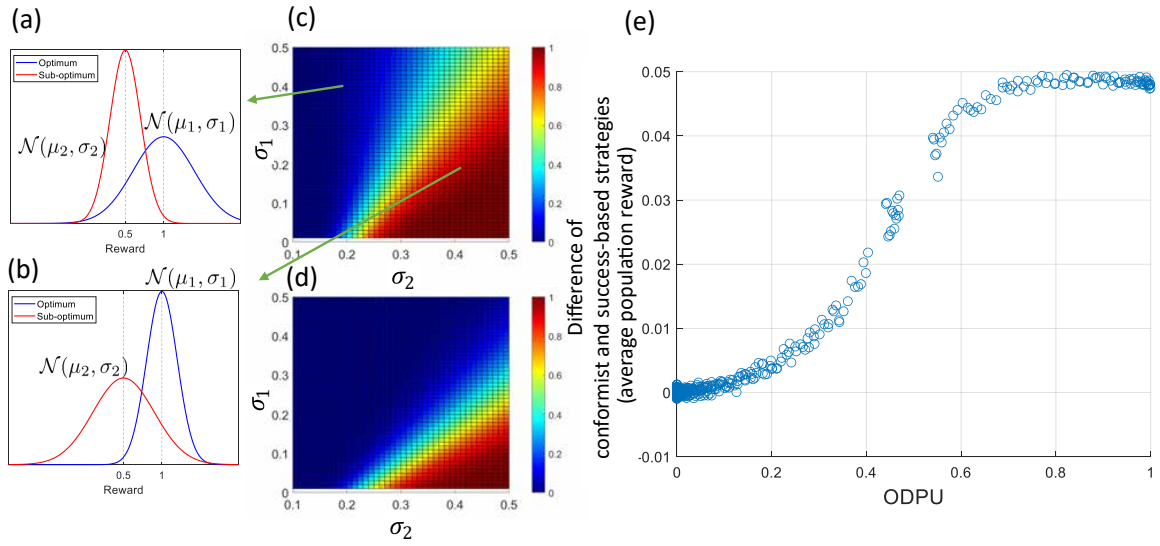


Figure 3: The higher the uncertainty between two distributions (measured by the ODPU) the higher the performance difference between conformist and success-based strategies in terms of average population reward at the end of simulation processes.
Figures (a) and (b) illustrate two cases where $\sigma_1 = 0.4$, $\sigma_2 = 0.2$ and $\sigma_1 = 0.2$, $\sigma_2 = 0.4$ respectively. (c) and (d) show the ODPU, formalized as the probability of sampling the highest reward value from the sub-optimum distribution, depending on σ_1 and σ_2 and the ratios of samples drawn independently from the optimum and sub-optimum reward distributions. In (c) the ratios of samples drawn from the optimum and sub-optimum distributions are 0.05 and 0.95, and in (d) the ratios are 0.5 and 0.5 respectively. (e) shows the relation between the ODPU and the difference in average reward at the end of the process between the populations with conformist and success-based strategies.

Figure 3e shows the strong correlation between the ODPU and the difference between performance of the conformist and of the success-based strategy (with Pearson's correlation coefficient $r = 0.9791$). Note that after a certain ODPU value (i.e. approximately around 0.1 – 0.2 in Figure 3e), their performance difference becomes highly significant. The maximum possible performance difference depends on the difference in their μ.

2.3 Meta-social learning hypothesis

In this section, we propose meta-social learning as a way for agents to arbitrate between individual and social learning strategies during their lifetime. We explored this hypothesis using several approaches. First, we used context encoding approach to determine the "context" of the environment by estimating three environmental variables, namely, environment change ($EC(t)$), conformity ($C(t)$) and uncertainty ($U(t)$) that play crucial role in the performance of the individual and social learning strategies. Then, based on our analysis, we defined meta-social learners that can arbitrate between these strategies depending on the context of the environment. Furthermore, as alternative

approaches, we used evolutionary algorithms and reinforcement learning to optimize the meta-social learners. Finally, we defined several baseline strategies that perform individual and social learning strategies randomly with a predefined probabilities.

2.3.1 Context encoding

We utilize the "social information" in estimating environment change, conformity and uncertainty. the social information is assumed be available for all individuals in the population and it consists of the action distribution of the population $h(a_j, t) \in \boldsymbol{H}$ and the rewards received by the individuals $r_i(a_j, t) \in \boldsymbol{R}$ where $h(a_j, t)$ denotes the frequency of action a_j in the population and $r_i(a_j, t)$ denotes the reward of individual i by performing action a_j at time t. From the reward distribution, it is possible to estimate average rewards $\mu'_j(t)$ and standard deviations $\sigma'_j(t)$ of actions a_j. Note that the social information is the same as the information[2] required to perform the success-based and conformist strategies.

Environment change. It is defined as the difference between the current and previous values of average rewards of the estimated optimum action:

$$EC(t) = \begin{cases} 1, & \text{if } |\mu'_*(t) - \mu'_*(t-\delta)| > th_{ec}, \\ 0, & \text{otherwise.} \end{cases}$$

where subscript $*$ denotes the estimated optimum action (that is the action with the highest average reward), δ is a parameter for comparing previous values of the average rewards, and th_{EC} is threshold for triggering the environment change detection. Threshold th_{EC} can depend on the task. In our experiments, we assign 0.15 for this threshold.

Conformity. It is based on the estimation whether the majority of the individuals are performing the behavior with the highest average reward. Thus, it is defined as:

$$C(t) = \begin{cases} 1 \text{ (conformity)}, & \text{if } \arg\max_j \mu'_j(t) = \arg\max_j h(a_j, t), \\ 0 \text{ (non-conformity)}, & \text{if } EC(t) = 1, \\ 0 \text{ (non-conformity)}, & \text{otherwise.} \end{cases}$$

Furthermore, if an environment change is detected, conformity is reset to 0.

Uncertainty. Estimated based on the ODPU (the probability of sampling higher reward values from sub-optimum distributions, see Sections 2.2 and 4.4). We use average rewards $\mu'_j(t)$ and

[2]Success-based social learning requires finding the action with the highest reward and conformist strategy requires finding the action with the highest frequency.

standard deviations $\sigma'_j(t)$ to compute the ODPU. Then, the uncertainty is detected as:

$$U(t) = \begin{cases} 1 \text{ (high uncertainty)}, & \text{if } ODPU > th_u, \\ 0 \text{ (low uncertainty)}, & \text{otherwise.} \end{cases}$$

where uncertainty threshold th_u is set to 0.1 in our experiments based on our uncertainty analysis in Section 2.2.

2.3.2 Meta-control of social learning strategies

A generic representation of the strategy selection process of a meta-social learner is shown in Equation 1. A strategy $S \in \{\textit{individual learning, success-based and conformist}\}$ is selected by meta-social learner $MSL()$ based on the context of the environment.

$$S := MSL(EC(t), C(t), U(t)) \tag{1}$$

In addition, we implement several other meta-social learning control mechanisms using various approaches and discuss under four types as follows (for implementation details of these algorithms, see Section 4.6):

Observation-based control. Here, we implement three versions. All of these versions start the process by using individual learning strategy. Similarly, they switch back to individual learning after an environment change. Otherwise, they use success-based or conformist social learning depending on the conformity and uncertainty.

- SL-EC-Conf (uses environment change and conformity) switches to the conformist social learning strategy if conformity ($C(t)$) is satisfied.
- SL-EC-Succ (uses environment change and uncertainty) switches to the success-based strategy in low uncertainty environment. Otherwise, they use individual learning.
- SL-EC-Conf-Unc (uses environment change, conformity and uncertainty) arbitrates between social learning strategies depending on conformity and uncertainty. If the conformity is observed in the population, then the individuals switch to the conformist strategy. Otherwise, if the environment is with low uncertainty, then they use the success-based strategy. If none of the above conditions is met, they perform individual learning.

Evolutionary control. The control policies for arbitrating between individual and social learning strategies were achieved by evolutionary algorithms. We used different environments (provided in Supplementary Material) for training and testing. We performed evolutionary-based training for 10 independent runs and selected the best performing strategy. Then, we tested this strategy on the test environments that were not encountered during the training processes and reported the results.

The goal of this separation was to demonstrate the generalization capability of the train model. We implemented two versions.

- SL-GA (trained by the genetic algorithms) perform the task based on the rules optimized with the genetic algorithms [16]. These rules are optimized to select individual and social learning strategies depending on the binary states of $EC(t)$, $C(t)$ and $U(t)$ (see Section 4.6.2). It is possible to explore all possible rules (based on all possible states of $EC(t)$, $C(t)$ and $U(t)$) in this space to identify the "optimal" rule. We note that the best performing evolutionary control mechanism[3] found by SL-GA converged to SL-EC-Conf-Unc.

- SL-NE (artificial neural network trained by neuroevolution) utilizes an artificial neural network to control meta-social learning, whose parameters were optimized by an evolutionary algorithm (known as neuroevolution approach [56]). The fully connected feedforward network (FCN) with one hidden layer takes the average rewards of arms $(\mu'_1(t), \ldots, \mu'_k(t))$, standard deviations $(\sigma'_1(t), \ldots, \sigma'_k(t))$ and the frequencies of the individuals that select k arms $(h(a_1, t), \ldots, h(a_k, t))$, and chooses a strategy as follows:

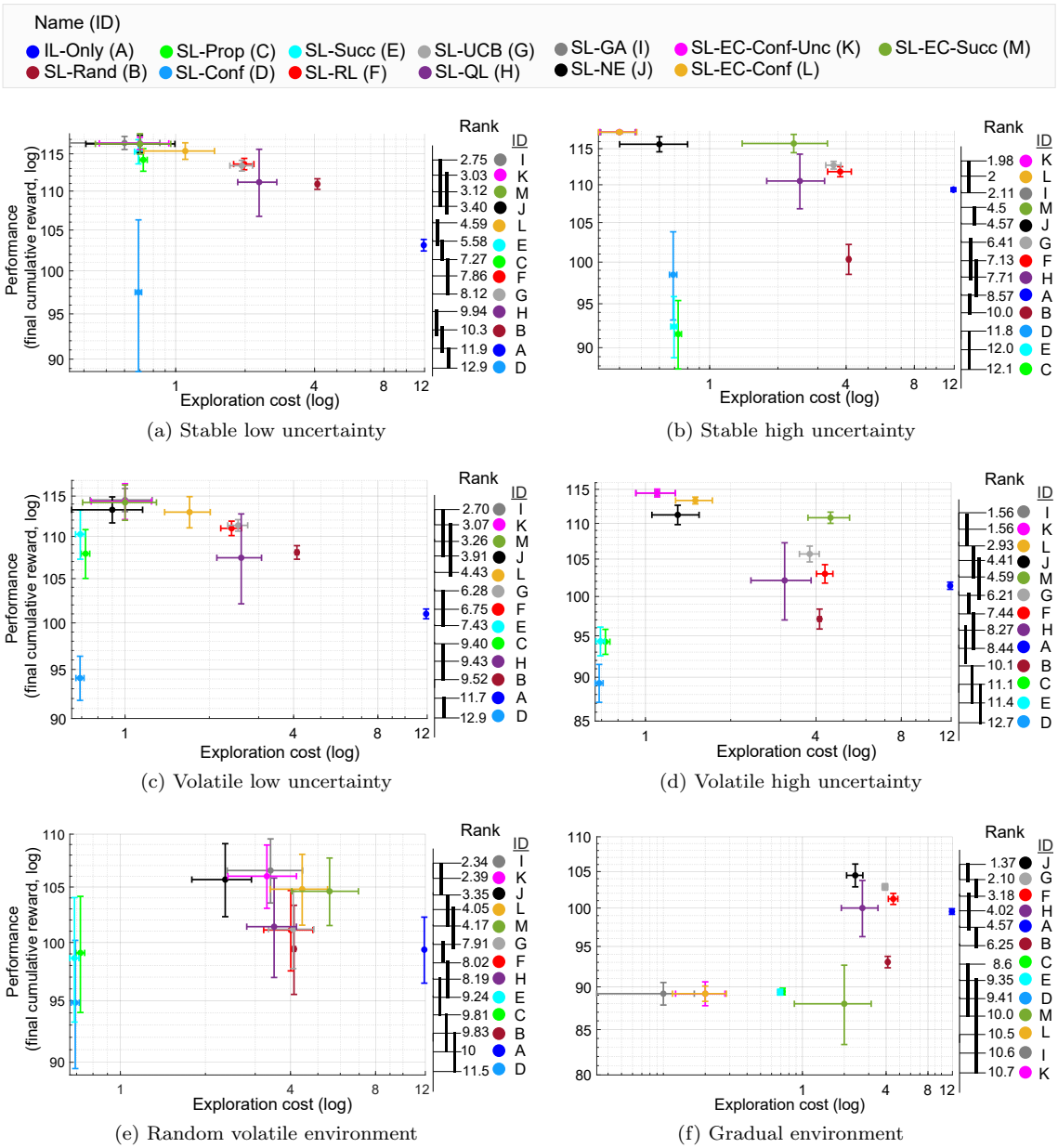
$$S := FCN(\mu'_1(t), \ldots, \mu'_k(t), \sigma'_1(t), \ldots, \sigma'_k(t), h(a_1, t), \ldots, h(a_k, t)) \tag{2}$$

  Note that unlike other versions of meta-social learning, this one does not require identifying the context of the environment such as environment change, uncertainty or conformity.

Multi-armed bandit control. We used SL-RL (ϵ-greedy algorithm), SL-UCB (upper confidence bound algorithm) and SL-QL (Q-learning) algorithms to learn to choose between individual, success-based and conformist social learning strategies. SL-RL and SL-UCB does not use environment context, rather, they perform a strategy selection based on the estimated rewards of selecting these strategies. The estimated rewards of these strategies are updated based on the rewards received after their selection. The SL-QL uses a value learning, intended to maximize the expected amount of future rewards. We used the context of the environment (environment change, uncertainty and conformity) as the states.

Other baseline strategies. We further implemented a set of baseline strategies: SL-Rand, SL-Prop, SL-Conf, SL-Succ, and IL-Only that perform a random strategy with a predefined fixed probability (see Section 4.6 for details). These meta-social learning strategies do not make use if the context of the environment.

[3]We perform 10 independent evolutionary runs and select the best control mechanism based on their cumulative reward. The results of the training processes are provided in Supplementary Material.



(a) Stable low uncertainty

(b) Stable high uncertainty

(c) Volatile low uncertainty

(d) Volatile high uncertainty

(e) Random volatile environment

(f) Gradual environment

Figure 4: SL-GA and SL-EC-Conf-Unc show the best performance vs. exploration cost on diverse set of environments. Overall, the meta-social learning strategies that utilize conformist strategy show better performance environments with high uncertainty.
On the right of each figure, the labels of the meta-social learning strategies (A through M) were ranked from best to worst based on their performance values. Decimal numbers on the left indicate their average ranks (the lower is the better), and the differences in their ranks that are not statistically significant at $\alpha = 0.05$ are linked by vertical black lines.

2.4 Uncertainty based meta-control resolves the trade-off between performance and exploration cost

To compare the performance of the meta-social learning algorithms on a task with uncertainty changes, first we defined stable and volatile environments with low and high uncertainty (See Supplementary Material, Section A.5). To construct these environments, we arbitrarily generated a set of six reward distributions (Figures 13a through 13f) from the highest to the lowest levels of uncertainty. Then, we created a task consisting of multiple periods, each of which is associated with a reward distribution selected from this set (**Experiment1**; Figures 15a, 15b, 15c and 15g). Moreover, we ran additional tests with two challenging tasks: random volatile environment where the number of environment changes and distributions were chosen randomly (**Experiment2**), and uncertain environment with a gradual environment change (**Experiment3**). The performance-exploration cost[4] trade-off and their ranks (based on CD diagrams) are shown in Figure 4. The change of the average reward and cumulative average reward during the learning processes on these environments are shown in Figure 15 in Supplementary Material A.5.

Overall, both the SL-GA and SL-EC-Conf-Unc achieved the highest performance with lowest exploration cost (the performance of the two models are not significantly different; Figure 4). This is due to the fact that the evolved controller in SL-GA is converged to the same controller used in the SL-EC-Conf-Unc. We note that the SL-NE provides one of the top five ranking results even though it uses low level population based features with artificial neural networks.

In general, the models utilizing the conformist strategy showed better performance in uncertain environments, compared to the ones with the success-based strategy. From the exploration cost point of view, we note that the IL-Only suffers from the highest cost with about three times more costly than the second most costly meta-social learner. In the case of the uncertain environment with gradual environment change (**Experiment3**; see Figure 4f), it is not surprising that the algorithms relying on the threshold-based environment change detection (SL-EC-Conf, SL-EC-Succ, SL-EC-Conf-Unc) did not perform well, which is ascribed to the failure in detecting environment change. To the contrary, SL-NE showed reliable performance robust against environment changes even though it was trained on the environments with different conditions. Note that this remarkable adaptation ability does not require an explicit environment change detection mechanism.

2.5 Uncertainty based meta-control dominates the evolution in volatile environments

What if there is a competition between different meta-social learning strategies in environments with various levels of volatility and uncertainty? Which ones would persist in the populations and become dominant relative to others? To assess that, we conducted an evolutionary analysis on meta-social

[4]Exploration is provided only through individual learners and controlled by ϵ, therefore, we measure the exploration cost by the number of individual learners used throughout a process multiplied by their ϵ.

learning. We further recorded how long they stay in the population (age) to assess their resilience. This analysis shows us successful strategies that are not being invaded by other strategies even in changing environmental conditions.

In the beginning of the evolutionary processes, we assigned each individual a specific type of meta-social learning strategy, randomly sampled from the complete set of the social learning strategies (i.e. IL-Only, SL-Rand, SL-Prop, ..., SL-EC-Conf-Unc). The individuals then used their own meta-social learning strategy to perform the tasks. After each generation, meta-social learners are selected based on the probability proportional to their fitness values (rewards received in the previous generation). Furthermore, we introduce a mutation operator that re-samples the type of meta-social learning strategy of each individual at each generation based on a small probability controlled by mutation rate mr. At every generation, the age of all strategies is increased by one. It is possible to pass multiple copies of a strategy to the next generation during the selection process. In this case, multiple copies are treated as offspring where only the age of the first copy is preserved while the age of the others is set to zero. Similarly, after mutation the age of the strategy is set to zero.

Figure 5 shows the population dynamics of meta-social learning algorithms during the evolutionary processes[5]. The meta-social learning strategies that perform well relative to the others show increase in their ratios in the population, whereas, the ones that cannot perform well show decrease in their ratios, and eventually die out[6].

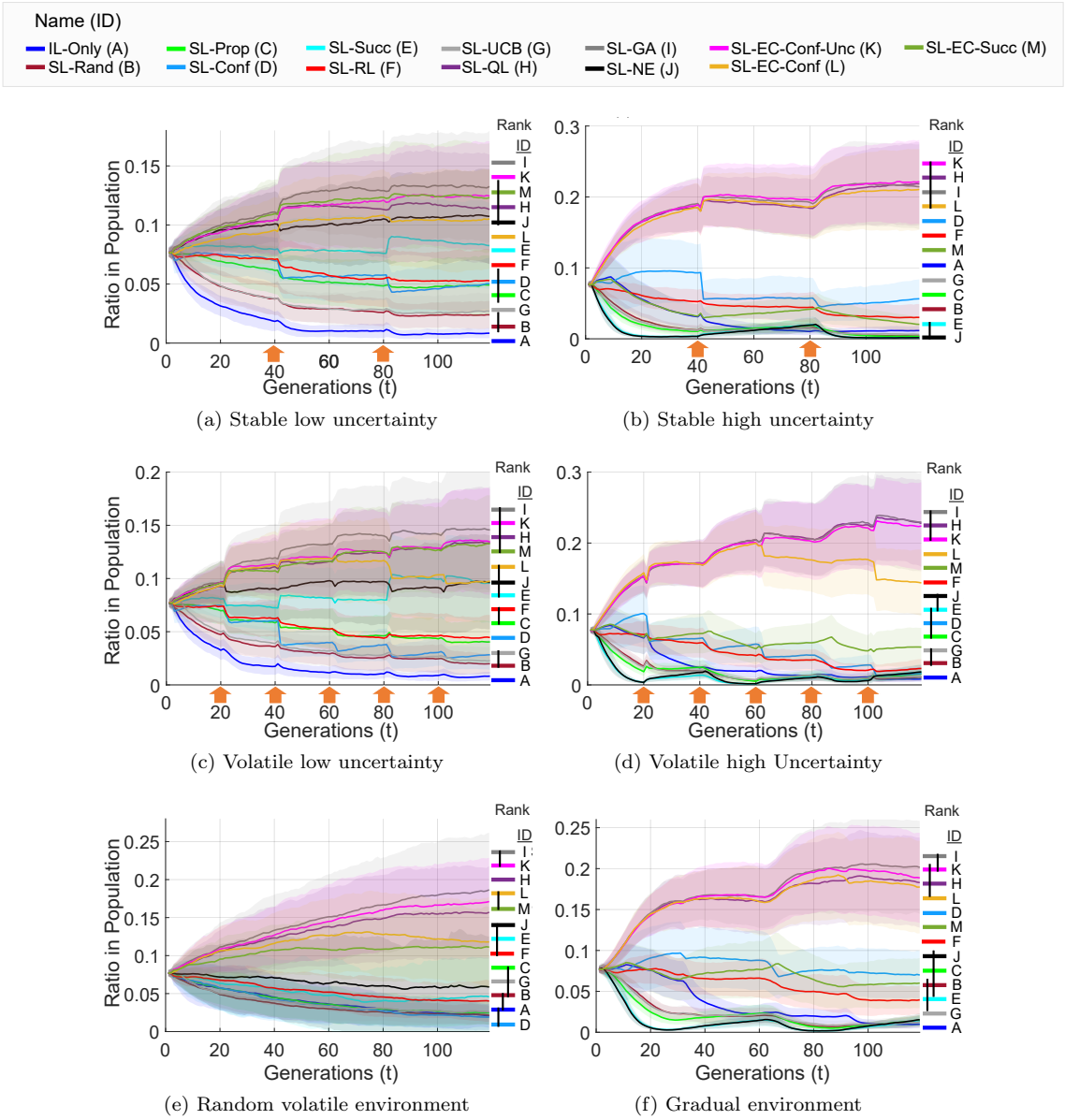
Overall, in the environments with low uncertainty, seven meta-social learning strategies show an increase in their ratios at the end of the processes relative to their starting ratios, whereas, in the environments with high uncertainty, only four of them show increase in their ratios. The most dominating four meta-social learners are: SL-EC-Conf-Unc, SL-GA, SL-EC-Conf and SL-QL. These meta-strategies, with the exception of SL-EC-Conf, make use of the environment uncertainty information. Remarkably, SL-EC-Conf is able to compete with the others using conformity bias without the need of using environment uncertainty.

We note that even though meta-social strategies such as SL-GA, SL-EC-Conf-Unc that failed to perform well in gradual environment (see Figure 4f) by themselves, they can show domination over other strategies in this experiment (see Figure 5f). This is due to the fact that the populations that consist of only these strategies cannot detect the environment change leading to the failure of exploring the other arm after an environment change. However, when they are used in combination with the other meta-social strategies (e.g. SL-Rand, SL-IL, SL-NE, ...), the exploration performed by other meta-social strategies helps overcome this effect. Consequently, these strategies (that fail due to the environment detection mechanisms) can perform well and dominate the populations.

We performed an analysis on the age distributions of the meta-social learning strategies during the evolutionary processes (provided in Figure 16 in Supplementary Material A.6). Sudden changes

[5]For statistical significance, we used a population consisting of 5000 individuals and perform the evolutionary processes for 112 independent runs. The statistical significance of the results measured by pairwise p-values and differences that are not statistically significant shown on the right side of each figure.

[6]Due to the use of mutation (with rate of 0.005), strategies are not completely eliminated from the evolutionary processes.



(a) Stable low uncertainty

(b) Stable high uncertainty

(c) Volatile low uncertainty

(d) Volatile high Uncertainty

(e) Random volatile environment

(f) Gradual environment

Figure 5: Based on the ratios in the populations, SL-GA, SL-EC-Conf-Unc, SL-QL and SL-EC-Conf are the most dominating meta-social learning strategies in wide range of environments.
On the right of each figure, the ranks of the algorithms (higher to lower in terms of ratios) at the end of the processes are shown. The differences that are not statistically significant (at $p > 0.05$, Wilcoxon rank-sum test) are linked using black vertical lines.
The points of environment change are indicated with orange arrows. The highlighted areas show the standard deviation 112 runs.

in the age distribution due to the environment change can be observed. The dominant meta-social learners show higher life expectancy throughout the processes especially in the environments with high uncertainty.

We further hypothesize that the life expectancy of the dominant meta-social learning strategies should be consistent with key variables of the evolutionary process, such as the selection strength and mutation rate. To test this we performed further experiments, in which we ran the same simulations while varying the level of selection strengths and mutation rates (as *low*, *moderate* and *high*). The selection strength is given by $p_i = f_i^s / \sum_{j=1}^{m} f_j^s$, where p_i is the selection probability of a meta-social learning strategy i, to pass to the next generation, f_i is its fitness value, and m is the total number of meta-social learners.

While the selection strength increases, the life expectancy of the dominant strategies increases, whereas, the life expectancy of the others decreases (provided in Supplementary Material A.7, see Figure 17). Furthermore, it can be observed that, while the mutation rate increases, the life expectancy of the dominant strategies decreases. This is due to the fact that, when the mutation rate is high, the probability of randomly mutating a dominant strategy increases, causing their life expectancy to decrease.

3 Discussion

While previous research examined individual and social learning strategies in the context of a changing environment [50, 22, 27, 1, 30, 29], this study tested a new hypothesis that measurements of environmental uncertainty can be used as a means to implement a reliable and cost-efficient learning strategy, regardless of environmental changes. To test this hypothesis, we performed an analysis on individual learning and two social learning strategies, namely success-based and conformist, on volatile and uncertain environments. Our analysis showed that the performance of the success-based strategy, the most direct way to explore to find the optimal policy, is susceptible to uncertainty in the environment, whereas that of the conformist strategy, though it does not guarantee the optimal performance, is highly reliable. Motivated by these results, we proposed several meta-social learning algorithms. Overall, the proposed meta-social learning strategies showed significantly better performance, and in the evolutionary analysis, they dominated other meta-social learning approaches in terms of survival rate.

The proposed meta-social learning scheme is motivated by the recent theoretical idea in neuroscience, suggesting that the brain uses meta-learning strategies to find a compromise between different types of learning, such as Pavlovian, model-free, and model-based learning [9, 61, 49, 48]. Accumulating evidence suggests that meta-learning allows individuals to resolve environmental uncertainty efficiently [37, 32, 36, 6, 7]. Critically, this view is fully consistent with our finding that meta-social learning strategy mitigates the adverse effect of environmental uncertainty on performance. Our computational framework can thus be used to examine neural mechanisms of meta-social

learning [47].

Our meta-social learning framework provides a means to examine complex population dynamics, thereby helping us better understand the fundamental nature of biological learning and decision making [41, 6, 8]. For example, our meta-social learning principle would provide theoretical insight into why animals or humans often copy others' behavior and why society needs to achieve conformity, especially in highly volatile situations. It would also be possible to examine how animal societies cope with environmental uncertainty and volatility. Moreover, the meta-social learning scheme can be extended to explain various types of ecological interactions, such as symbiosis, mimicry, and mutualism.

In artificial intelligence and robotics applications, nature inspired approaches have proven to be successful in modeling intelligent behavior [5, 15, 44, 18].

Accordingly, social learning aims to benefit from the collective property of multi-agent systems to provide efficient learning and adaptation as a population. As illustrated in this work, exploiting the behaviors of other individuals can reduce exploration cost significantly. It can also improve learning efficiency in uncertain and volatile environments. This may prove to be important in real-world applications such as the internet of things and swarm robotics [66, 2, 40, 51, 14]. For instance, one recent example of distributed learning approach has been used in healthcare to detect illnesses [62]. Meta-social learning strategies can play a key role in these kinds of distributed learning applications to improve the efficiency of learning.

Rational choice theory in economics and game theory suggests that individuals choose their best action through a cost-benefit analysis which we usually conceptualize as involving explicit deduction (thinking through pros and cons) [52, 54]. Since our results suggest that an individual can make cost-effective decisions instead via social information, i.e., the decisions of others and their outcomes, it may be useful to consider models based on such foundations as well. For instance, it would be possible to estimate the environment uncertainty and volatility simply by measuring individual choice variability. This inference based on social information improves sample efficiency significantly compared to individual learning.

In addition to the computational and theoretical implications of meta-social learning, another exciting research direction is to use meta-social learning to examine the fundamental nature of social networks [11]. For instance, complex dynamics of individual interactions can lead to the emergence of various "social learning networks". Investigation of fundamental computations underlying the emergence and evolution of social networks would allow us to understand and predict the future of animal societies. Furthermore, the computational framework of meta-social learning be used to test new hypotheses about multi-agent social learning [38, 3]. For instance, it is possible to test whether complex internal dynamics arising from meta-social learning promote the natural emergence of curriculum.

4 Methods

4.1 Multi-armed bandit problem

The multi-armed bandit problem is a classic problem in reinforcement learning that [50, 34, 50, 58] where individuals are required to perform actions to choose one of k alternative choices (also known as k-arms). Performed actions provide rewards based on their underlying distributions that is unknown to the individual. In our work, We model the reward distribution of each action as a Gaussian distribution $\{\mathcal{N}(\mu_1, \sigma_1), \ldots, \mathcal{N}(\mu_k, \sigma_k)\}$. The goal of an individual is to perform actions to choose repetitively one of the choices and collect the rewards for a certain period of time such a way that can maximize the cumulative sum of the rewards received during the process.

4.2 Individual learning model

Individual learning is modeled as a reinforcement learning agent using ϵ-greedy algorithm [58]. We denote estimated reward of an action a at time t as $Q(a, t)$. The reward estimation is updated based on the rewards $r(a, t)$ when a is chosen using Equation (3).

$$Q(a, t+1) = Q(a, t) + \beta \left[r(a, t) - Q(a, t) \right] \tag{3}$$

where $0 < \beta \leq 1$ is the step size which was set to 0.2 in our experiments. Equation (4) shows the ϵ-greedy algorithm where the behavior with the highest estimated reward or a random behavior is chosen with the probabilities of $1 - \epsilon$ and ϵ respectively. We set $\epsilon = 0.1$ in our experiments.

$$a(t) = \begin{cases} \arg\max_a Q(a, t), & \text{based on the probability } 1 - \epsilon, \\ \text{random}(a), & \text{based on the probability } \epsilon. \end{cases} \tag{4}$$

4.3 Social learning models

Social learners copy the behavior of other individuals in the population based on a certain strategy [31]. We implement two social learning strategies as: success-based and conformist given in below:

$$a(t) = \begin{cases} \arg\max_a r(a, t-\tau), & \text{if success-based strategy,} \\ \arg\max_a h(a, t-\tau), & \text{if conformist strategy.} \end{cases}$$

where $r(a, t-\tau)$ and $h(a, t-\tau)$ denote the reward and frequency of an action a at time $t-\tau$ with some latency τ. Social learning is performed only when $t - \tau > 0$.

In success-based strategy, social learners copy the behavior of the individual with the best reward at $t-\tau$, and in conformist strategy, social learners copy the most frequent behavior in the population at time $t - \tau$.

4.4 Optimum distribution prediction uncertainty

Uncertainty of the environment ($U(t)$) is estimated based on the probability of sampling higher reward values from sub-optimum distributions. We refer to this probability as the optimum distribution prediction uncertainty (ODPU) and define as:

Let $\{X^1_{j_1}\}_{j_1}, \ldots, \{X^k_{j_k}\}_{j_k}$ be k sets of normally distributed random variables, where the random variables of set i are independently drawn from $\mathcal{N}(\mu_i, \sigma_i)$. All k sets are finite, where j_i represents an integer value such that $0 < j_i \leq N_i < \infty$ for all $i = 1, \ldots, k$. The notation $X^i_{(N_i)}$ is used to indicate the N_i-th order statistic. That is, the maximum of all random variables of a given set $\{X^i_{j_i}\}_{j_i}$.

Now, using the fact that the random variables are independently drawn from normal distributions, one can write the probability density function, f, and the cumulative distribution function, F, of the $N_{(i)}$-th order statistic as

$$f_{X^i_{(N_i)}}(x) = \frac{N_i}{\sigma_i}\phi\left(\frac{x-\mu_i}{\sigma_i}\right)\Phi\left(\frac{x-\mu_i}{\sigma_i}\right)^{N_i-1}, \qquad F_{X^i_{(N_i)}}(x) = \Phi\left(\frac{x-\mu_i}{\sigma_i}\right)^{N_i}.$$

The optimum distribution prediction uncertainty is then be formulated by

$$\begin{aligned}\text{ODPU} &= 1 - \mathbb{P}(X^1_{(N_1)} \geq X^i_{(N_i)}; \quad \text{for all} \quad i \geq 2),\\ &= 1 - \int_{-\infty}^{\infty} f_{X^1_{(N_1)}}(y) \cdot \left(F_{X^2_{(N_2)}}(y) \cdot \cdots \cdot F_{X^k_{(N_k)}}(y)\right) \mathrm{d}y,\end{aligned}$$

which describes the probability of sampling higher values from the distributions with lower means relative to μ_1.

4.5 Evolution of social learning

4.5.1 Mathematical model

This section provides our mathematical model for analysing the evolution of social learning strategies on the multi-armed bandit problem for number of arms $k = 2$. Let $a_i, i \in \{1, 2\}$ denote actions with corresponding payoffs $r(a_i, t)$ at time t.

The frequency of the individual learners in the population at time t is denoted as $IL(t)$. We further distinguish two types of individual learners $A_1(t)$ and $A_2(t)$ to indicate the frequencies of the individuals that perform actions a_1 and a_2. The sum of the frequencies of the behaviors satisfy the following condition: $A_1(t) + A_2(t) = IL(t)$ for all $t \geq 0$.

We denote the frequencies of the social learners in the population as $SL(t)$ at time t. Overall, the sum of the individual and social learners in the population satisfy the following condition: $IL(t) + SL(t) = 1$ for all $t \geq 0$.

The change in the frequencies of A_1, A_2 and SL are modeled using the replicator-mutator

equation [33, 45] given as a system of coupled first order ordinary differential equations below [7]:

$$\begin{cases} \dot{A}_1(t) = \mathbf{F}(t)[\mathbf{I}^T(t) \circ \text{col}_1(\mathbf{M})] - A_1(t)\psi(t), & t > 0, \\ \dot{A}_2(t) = \mathbf{F}(t)[\mathbf{I}^T(t) \circ \text{col}_2(\mathbf{M})] - A_2(t)\psi(t), & t > 0, \\ \dot{SL}(t) = \mathbf{F}(t)[\mathbf{I}^T(t) \circ \text{col}_3(\mathbf{M})] - SL(t)\psi(t), & t > 0, \\ A_1(t) = A_{1,0}, A_2(t) = A_{2,0}, SL(t) = SL_0, & t = 0. \end{cases} \tag{5}$$

where $\mathbf{F}(t) := [f_{A_1}(t), f_{A_2}(t), f_{SL}(t)]$ is a row vector of fitness values, $\mathbf{I}(t) := [A_1(t), A_2(t), SL(t)]$ is a row vector of individual frequencies, $\circ$ denotes the element-wise multiplication operator, $\text{col}_k()$ is a function that returns the k-th column of a matrix, and $\psi(t)$ is the average fitness of the population found as:

$$\psi(t) := \mathbf{F}(t)\mathbf{I}^T(t) \tag{6}$$

The replication may not be perfect. The mutation probabilities are provided by $\mathbf{M}$ where M_{ij} indicates the probability that type j is produced by type i, and $\mathbf{M}_{i:}$ indicates row vector with an index of i. $\mathbf{M}$ is a row-stochastic matrix thus satisfies the following condition:

$$\mathbf{M} \in \{\mathbf{A} \in \mathbb{R}^{3\times 3}_{\geq 0} \colon \sum_{j=1}^{3} A_{ij} = 1,\ 1 \leq i \leq 3\}.$$

In our experiments, we set mutation matrix $\boldsymbol{M}$ as follows:

$$\boldsymbol{M} = \begin{bmatrix} 0.995 & 0 & 0.005 \\ 0 & 0.995 & 0.005 \\ 0.0025 & 0.0025 & 0.995 \end{bmatrix}$$

which indicates mutation rate of 0.005 from the individual learners to social learner and vice versa.

Individual learners perform actions a_1 and a_2. However, they try the other action with a small frequency ϵ for exploration (e.g. analogous to ϵ-greedy algorithm in reinforcement learning [58]). Thus they suffer from an exploratory cost. On the other hand, this may become useful for learning new action in case if the environment changes (i.e. change in the payoffs of the actions). Consequently, the fitness f_{A_i} of type A_i is found by the weighted average of payoffs obtained from performing different actions as shown in Equation 7.

$$f_{A_i}(t) = (1 - \epsilon)r(A_i, t) + \epsilon r(A_j, t) \tag{7}$$

for all $i, j = \{1, 2\}$ where $i \neq j$.

Fitness of the the social learners $f_{SL}(t)$ updated based on a specific social learning strategy. We define four SLSs in following sections.

[7] Dot notation represents time derivative (i.e. $\dot{x} = dx/dt$).

Success-based. the social learners copy the behavior of successful individual from a previous time $(t-\tau)$. Thus, the fitness values of the social learners equal to the reward received by performing the optimum action.

$$f_{SL}(t) = r(a^*, t) \tag{8}$$

where a^* denotes the optimum action $r(a^*, t)$ is its reward at time t.

Conformist. the social learners copy the behavior with the highest frequency in the population. We keep track of the behavior frequencies in the population and introduce some latency represented as τ. The frequencies of the behaviors performed by the social learners are also included into the model. We first show how the frequencies of the behaviors are computed and then define the fitness of conformist strategy.

Let $h(a_i, t)$ denote the frequencies of actions a_i at time t. Furthermore, we define $H_{SL}(a_i, t)$ to denote the frequencies of the actions performed by the social learners. When $t = 0$, some portion $(H_{SL}(0))$ of the population consists of social learners, however they do not perform any action because the information of the frequencies of the actions in previous times $(t - \tau)$ is not available. Therefore, we set the initial values when $(t - \tau <= 0)$ as $H_{SL}(a_i, t) = 0$, $h(a_n, 0) = A_i(0)$, and $f_{SL}(0) = 0$.

When $t > \tau$, we update the frequencies of the actions as follows:

$$H_{SL}(a_i, t) = \begin{cases} 1, & \text{if } i = \arg\max_i h(a_i, t - \tau), \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

$$h(a_i, t) = A_i(t) + SL(t) h_{SL}(a_i, t) \tag{10}$$

Finally, the fitness of social learners, given in Equation (11), is found by the average payoffs of the social learners that perform each behavior type.

$$f_{SL}(t) = \sum_{i=1}^{n} h_{SL}(a_i, t) r(a_i, t). \tag{11}$$

4.5.2 Evolutionary algorithm

Algorithm 1 provides the pseudocode for the evolutionary algorithm [16] we use to analyse the evolution of a population of individual and social learners [27, 28]. Individuals are assigned one of these types randomly during the initialization process. In each generation, the individuals can perform their actions based on their type. Their fitness is computed based on their actions and used for the selection process for the next generation. We use only a mutation operator which alters the type of a selected individual with a small probability.

The individual learners are modeled as reinforcement learning agents where they perform their behavior based on their model. Moreover, their models can be updated based on the rewards received as response to their behaviors. We use ϵ-greedy algorithm, discussed in detail in Section 4.2 [58], to

Algorithm 1 the Evolutionary Algorithm for the evolution of social learning strategies.

procedure EA(ϵ, sls) ▷ Evolution of individual and *sls* type social learners
  // ϵ: exploration parameter of individual learning
  // sls: the type of social learning strategy (i.e. success-based or conformist)
  $t = 1$ ▷ Generation counter t
  $mr := mutationRate$
  $I_t :=$ initializeIndividuals() ▷ Initial population
  while $t \leq T$ **do**
    for each $i \in I_t$ **do**
      if (isIndividualLearner(i) **or** $t = 1$) **then** ▷ Individual Learning
        $r_i(a_j, t) =$ individualLearning(ϵ)
        updateDecisionModel($r_i(a_j, t)$) ▷ See Equation (3)
      else ▷ Social Learning
        $r_i(a_j, t) =$ socialLearning(sls)
        updateDecisionModel($r_i(a_j, t)$)
      end if
      $f_i =$ updateFitness($r_i(a_j, t)$)
    end for
    $I' :=$ select(I_t, F)
    $I_{t+1} :=$ mutate(I', mr)
    $t = t + 1$
  end while
end procedure

model the learning process of the individual learners.

The social learners on the other hand, perform their behaviors based on the behaviors of others. We model the same strategies, namely, conformist and success-based discussed in Section 4.5.1. In case of conformist strategy, the social learners select the behavior with maximum frequency in the population. For the success-based strategy, social learners copy the behavior of the individual with the best fitness value in the previous generation.

Depending on the outcome $r_i(a_j, t)$, that is the reward received by the i-th individual performing action a_j, their fitness values $f_i \in F$ are updated. We simply use the reward as the fitness of an individual i at generation t as: $f_i(t) = r_i(a_j, t)$. The average population reward $\psi(t)$ is the average of the fitness values of the individuals in the population: $\psi(t) = \frac{1}{m} \sum_i^m f_i(t)$ where m is the number of individuals in the population.

Selection of the individuals for the next generation $t + 1$ is performed based on their fitness values at t. We use the *roulette wheel selection* (also known as *fitness proportionate selection* [16]) to simulate natural selection process. According to this scheme, m number of individuals are selected based on the probability that is proportional to their fitness values as given below:

$$p_i(t) = \frac{f_i(t)}{\sum_{j=1}^m f_j(t)}, \tag{12}$$

where $p_i(t)$ is the probability of selecting individual i from the population, and $f_i(t)$ is the fitness of the individual. The same individuals can be selected multiple times to construct the population for the next generation. There is no mating process involved. However, individuals are mutated by changing their type with a small probability controlled by the mutation rate (mr).

4.6 Meta-social learning

The meta-social learners can switch between individual and social learning during their lifetime. A generic algorithm for meta-social learning is provided in Algorithm 2. In this section, we provide the implementation details of all the algorithm variants used to control the meta-social learning strategies.

Algorithm 2 Meta-social learning algorithm based on the environmental variables: environment change, uncertainty and conformity.

procedure MSL(ϵ) ▷ Run independently for each individual
    // ϵ: exploration parameter of individual learning
    // S: type of learning strategy $\in$ {individual learning, success-based or conformist strategy}
    // t: discrete time counter
    $t = 1$ ▷ Initial t
    while $t \leq T$ **do**
        $[EC(t), C(t), U(t)]$:= ContextEncoding($\boldsymbol{H}, \boldsymbol{R}$) ▷ Estimate the context of the environmental
        S := MSL($EC(t), C(t), U(t)$) ▷ Meta-social learner
        if S = "individual learning" **then** ▷ Individual Learning
            $r_i(a_j, t)$ = individualLearning(ϵ)
            updateDecisionModel($r_i(a_j, t)$) ▷ See Equation (3)
        else ▷ Social Learning
            $r_i(a_j, t)$ = socialLearning(S)
            updateDecisionModel($r_i(a_j, t)$)
        end if
        $t = t + 1$
    end while
end procedure

4.6.1 Observation-based control

The SL-EC-Conf-Unc algorithm uses three environmental variables, environment change ($EC(t)$), conformity ($C(t)$) and uncertainty ($U(t)$), to decide the type of learning strategy. Initially, and after an environment change, the algorithm uses individual learning strategy for exploration.

During individual learning, the conformity of the population and uncertainty of the environment are estimated and used for switching between conformity and success-based strategies as shown in Table 1.

Table 1: The strategies implemented by the SL-EC-Conf-Unc based on the conformity and uncertainty.

	Conformity ($C(t) = 1$)	**Non-conformity** ($C(t) = 0$)
Low uncertainty ($U(t) = 0$)	Conformist	Success-based
High uncertainty ($U(t) = 1$)	Conformist	Individual learning

Other variants in this class include SL-EC-Conf and SL-EC-Unc. These two variants use environment change but does not make use off full functionality of SL-EC-Conf-Unc. In case of SL-EC-Conf, individuals can use only individual learning and conformist strategies depending on the environment change and conformity, whereas, in case of SL-EC-Unc, they can use only individual and success-base strategies based on the environment change and uncertainty.

4.6.2 Evolutionary control

SL-GA (trained by the genetic algorithms) uses genetic algorithms (GAs [16]) to optimize the meta-social learning policies of the individuals for switching between the individual and social learning strategies. Shown in Table 2, we encode the type of strategy s_j (i.e. individual learning, success-based or conformist) for a given state of the environmental variables, namely environment change, conformity and uncertainty. Since these variables can take binary values (see Section 4.6), there are 8 possible states (discrete variable) which can take three strategies. Thus, there are total of $3^8 = 6561$ possible distinct policies.

In addition to these discrete variables, we include two continuous variables into the genotype of individuals for determining the thresholds used in environment change (th_{ec}) and uncertainty (th_u). Consequently, the genotype of the individuals consist of 10 genes, 8 discrete and 2 continuous variables.

Table 2: Possible states of the environment and their strategy assignments that can take one of the strategies as: individual learning, success-based or conformist.

Environment Change	**Conformity**	**Uncertainty**	**Strategy**
0	0	0	s_1
0	0	1	s_2
...	...	...	...
1	1	1	s_8

We use a standard GA with population size of 50 individuals, roulette wheel selection with four elites, 1-point crossover operator with 0.8 probability and a mutation operator which selects discrete genes with the probability of $1/(L-2)$ where $L-2$ is the length of the genotype excluding the genes that encode continuous variables, replaces one of three possibilities for the discrete genes, and performs Gaussian perturbation with zero mean and 0.1 standard deviation on the continuous genes.

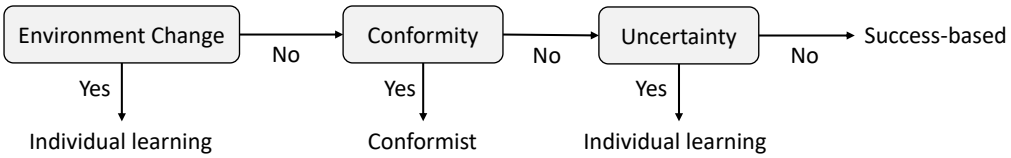


Figure 6: Discrete part of the SL-GA policy that achieved the highest cumulative reward. "Yes" and "No" indicate 1 and 0 states of the environmental variables shown in Table 2. Thresholds of the evolved rule for uncertainty and environment change is $th_u = 0.05$ and $th_{ec} = 0.15$.

We use a separate training environment for optimizing the meta-social learning policies using the GA (provided in Supplementary Material). The GA aims to maximize the fitness values of the policies which is computed by the median of the cumulative sum of the average population reward of 112 runs as follows:

$$f = \text{median}\left(\sum_{t=1}^{T} \psi_i(t)\right), \forall i = \{1, \ldots, 112\} \tag{13}$$

The GA process on the training environment is executed for 10 independent runs. We stop the evolutionary process if the algorithm fails to find a better fitness value for 20 subsequent generations. At the end of the runs, we select the best policy to be tested on the test environment reported in Results section.

Figure 6 shows the best evolved policy over 10 independent GA runs. The details of the optimization process provided in Supplementary Material. The evolutionary process is performed on a separate environment different than the environments we used for test in Results section. Note that the best evolved policy converged to the policy suggested by our analysis, and implemented by the SL-EC-Conf-Unc.

SL-NE (ANN based trained by neuroevolution) uses neuroevolution (NE) [56] approach to optimize artificial neural network (ANN) based policies. In NE, evolutionary algorithms are used for the optimization processes of the topologies and/or weights of the networks.

Illustrated in Figure 7, we use feed-forward ANNs with one hidden layer to perform individual learning, success-based and conformist social learning strategies. The input to the networks are the average and standard deviations of the estimated rewards of the actions, and frequencies of the individuals that perform each action (see Equation (2)). For two actions, we used $6, 12$ and 3 neurons in the input, hidden and output layers respectively. We include an additional bias neuron (constant $+1$) in input and hidden layers. Therefore, the total number of network parameters is $12(6+1) + 3(12+1) = 123$.

We use genetic and differential evolution (DE) algorithms to optimize the weights of the networks by directly mapping them into the genotype of the individuals and representing them as real valued vectors. In both algorithms, we use a population of 50 individuals, and initialize the weights of the first generation randomly from the uniform distribution from range $[-1, 1]$. In case of the GA,

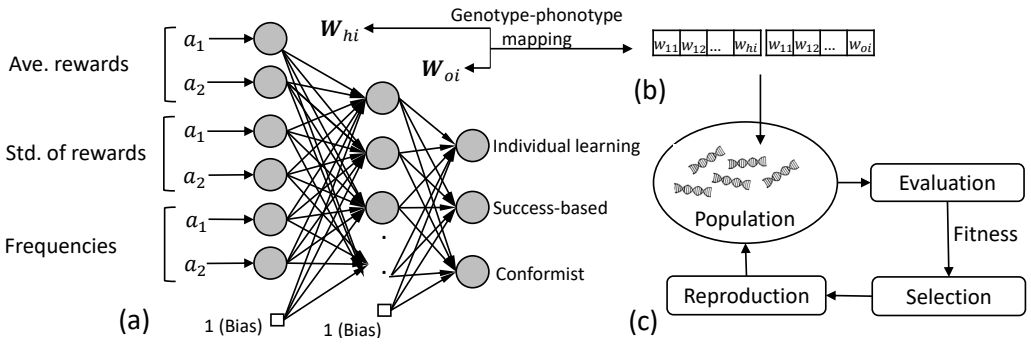


Figure 7: Neuroevolution scheme used to optimize the social learning policies. (a) Feed-forward artificial neural network topology with one hidden layer can take the average, standard deviations and frequencies of two actions a_1 and a_2 and decides to perform individual learning, success-based or conformist social learning strategies. (b) The weights of the networks between input and hidden layers (W_{hi}), and hidden and output layers (W_{oi}) are directly mapped to the genotype of the individuals and represented as real valued vectors. (c) Evolutionary algorithms are used to optimize the genotype of the individuals.

we use roulette wheel selection with 5 elites, 1-point crossover operator with the probability of 0.8 and Gaussian mutation operator as: $\mathcal{N}(0, 0.1)$, that performs independent perturbation for each dimensions in the genotype.

In the case of the DE, we use "rand/1" mutation strategy and uniform crossover with parameters of $F = 0.5$ and $CR = 0.1$ respectively [67, 57].

Both algorithms aim to maximize the median of the total rewards, given in Equation (13), on the training environment provided in Supplementary Material. We run the GA and DE for 10 independent runs each, and use the ANN that achieved the best fitness value for the comparison in Results section.

4.6.3 Multi-armed bandit control

1. SL-RL (ϵ-greedy algorithm): the action-value based approach used in individual learning model (discussed in Section 4.2) is used for selecting the social learning strategy. In this case, the action space consists of performing one of the followings: individual learning, success-based and conformist strategies at time t. $Q(a, t)$ is the estimate reward of these learning approaches and updated based on the rewards received as shown in Equation (3). One of the actions are selected based on Equation (4).

2. SL-QL (Q-learning): we represent the estimate rewards of a certain action a in a certain state of the environment s as $Q(s_t, a_t)$ at time t. An action is selected based on the following:

$$a(t) = \begin{cases} \arg\max_a Q(s_t, a), & \text{based on the probability } 1 - \epsilon_{QL}, \\ \text{random}(a), & \text{based on the probability } \epsilon_{QL}. \end{cases}$$

where ϵ_{QL} is exploration parameter for the Q-learning. Then, estimate rewards are updated based on the Bellman equation as follows:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha \left(r(a, t) + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right)$$

where α and γ are the learning rate and the discount factor.

Here, we use there environmental variables as states as: environment change $EC(t)$, conformity $C(t)$ and uncertainty $U(t)$. All of these variables are binary, thus, there are a total of eight states. There are three possible actions as: individual learning, success-based or conformist social learning strategies. We trained the SL-QL on the training environment provided in Supplementary Material. We performed experiments with different parameter settings of the algorithm and found that $\epsilon_{QL} = 0.2$, $\alpha = 0.01$ and $\gamma = 0$ value assignments provided the best results.

3. SL-UCB (upper confidence bound): to control the degree of the exploration, the equation for selecting actions is modified as follows:

$$a(t) = \arg\max_a \left[Q(a, t) + c \sqrt{\frac{\ln t}{N(a, t)}} \right]$$

where c is exploration parameter and $N(a, t)$ is the number of time a is selected until time t. We use the same update rule for $Q(a, t)$ given in Equation (3).

In UCB selection, the square root term is the uncertainty in the estimate of a. While $N(a, t)$ increases, uncertainty terms decreases, whereas, while $N(a, t)$ keeps the same, uncertainty increases (since t increases) [58].

4.6.4 Other baseline strategies

1. IL-Only (individual learning): individuals perform only individual learning throughout the processes.

2. SL-Rand (random strategies): individuals perform randomly one of individual learning, success-based and conformist social learning strategies with equal probability.

3. SL-Prop (proportional strategy selection): individuals perform success-based, conformist, individual learning strategies with probabilities of 0.45, 0.45 and 0.1 respectively.

4. SL-Conf (conformist with individual learning): individuals perform conformist and individual learning strategies with probabilities of 0.95 and 0.05 respectively.

5. SL-Succ (success-based with individual learning): individuals perform success-based and individual learning strategies with probabilities of 0.95 and 0.05 respectively.

Acknowledgement. This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government (MSIT) (No.2019-0-01371, Development of brain-inspired AI with human-like intelligence) and Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TC1603-52.

References

[1] Kenichi Aoki, JoeYuichiro Wakano, and MarcusW Feldman. The emergence of social learning in a temporally changing environment: a theoretical model. *Current Anthropology*, 46(2):334–340, 2005.

[2] Luigi Atzori, Antonio Iera, and Giacomo Morabito. The internet of things: A survey. *Computer networks*, 54(15):2787–2805, 2010.

[3] Bowen Baker, Ingmar Kanitscheider, Todor M. Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[4] Robert Boyd and Peter J Richerson. *Culture and the evolutionary process*. University of Chicago press, 1988.

[5] Nicolas Bredeche, Evert Haasdijk, and Abraham Prieto. Embodied evolution in collective robotics: A review. *Frontiers in Robotics and AI*, 5:12, 2018.

[6] Caroline J Charpentier, Kiyohito Iigaya, and John P O'Doherty. A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning. *Neuron*, 2020.

[7] Sven Collette, Wolfgang M Pauli, Peter Bossaerts, and John O'Doherty. Neural computations underlying inverse reinforcement learning in the human brain. *Elife*, 6:e29718, 2017.

[8] Isabelle Coolen, Y Van Bergen, Rachel L Day, and Kevin N Laland. Species difference in adaptive use of public information in sticklebacks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1531):2413–2419, 2003.

[9] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711, 2005.

[10] Lewis G Dean, Gill L Vale, Kevin N Laland, Emma Flynn, and Rachel L Kendal. Human cumulative culture: a comparative perspective. *Biological Reviews*, 89(2):284–301, 2014.

[11] Alain Degenne and Michel Forsé. *Introducing social networks*. Sage, 1999.

[12] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[13] Kaleda Krebs Denton, Yoav Ram, Uri Liberman, and Marcus W Feldman. Cultural evolution of conformity and anticonformity. *Proceedings of the National Academy of Sciences*, 2020.

[14] Julia T Ebert, Melvin Gauci, and Radhika Nagpal. Multi-feature collective decision making in robot swarms. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1711–1719, 2018.

[15] Agoston E Eiben, Evert Haasdijk, and Nicolas Bredeche. Embodied, on-line, on-board evolution for autonomous robotics. In *Symbiotic Multi-Robot Organisms: Reliability, Adaptability, Evolution*, pages 361–382. Springer Verlag, 2010.

[16] Agoston E Eiben, James E Smith, et al. *Introduction to evolutionary computing*, volume 53. Springer, 2003.

[17] Ben Eppinger, Thomas Goschke, and Sebastian Musslick. Meta-control: From psychology to computational neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, pages 1–6, 2021.

[18] Evert Haasdijk, Nicolas Bredeche and Agoston E Eiben. Combining environment-driven adaptation and task-driven optimisation in evolutionary robotics. *PloS ONE*, 9(6), 2014.

[19] Liane Gabora. An evolutionary framework for cultural change: Selectionism versus communal exchange. *Physics of Life Reviews*, 10(2):117–145, 2013.

[20] Jan Gläscher, Nathaniel Daw, Peter Dayan, and John P O'Doherty. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595, 2010.

[21] Roderich Groß, Alasdair I Houston, Edmund J Collins, John M McNamara, François-Xavier Dechaume-Moncharmont, and Nigel R Franks. Simple learning rules to cope with changing environments. *Journal of the Royal Society Interface*, 5(27):1193–1202, 2008.

[22] Joe Henrich and Robert Boyd. The evolution of conformist transmission and the emergence of between-group differences. *Evolution and human behavior*, 19(4):215–241, 1998.

[23] Joseph Henrich. *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press, 2017.

[24] Cecilia Heyes. Who knows? metacognitive social learning strategies. *Trends in cognitive sciences*, 20(3):204–213, 2016.

[25] Cecilia Heyes. When does social learning become cultural learning? *Developmental Science*, 20(2):e12350, 2017.

[26] Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American mathematical society*, 40(4):479–519, 2003.

[27] Tatsuya Kameda and Daisuke Nakanishi. Cost–benefit analysis of social/cultural learning in a nonstationary uncertain environment: An evolutionary simulation and an experiment with human subjects. *Evolution and Human Behavior*, 23(5):373–393, 2002.

[28] Tatsuya Kameda and Daisuke Nakanishi. Does social/cultural learning increase human adaptability?: Rogers's question revisited. *Evolution and Human Behavior*, 24(4):242–260, 2003.

[29] Anne Kandler and Kevin N Laland. Tradeoffs between the strength of conformity and number of conformists in variable environments. *Journal of theoretical biology*, 332:191–202, 2013.

[30] Jeremy Kendal, Luc-Alain Giraldeau, and Kevin Laland. The evolution of social learning rules: payoff-biased and frequency-dependent biased transmission. *Journal of theoretical biology*, 260(2):210–219, 2009.

[31] 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.

[32] Dongjae Kim, Geon Yeong Park, PO John, Sang Wan Lee, et al. Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. *Nature communications*, 10(1):1–14, 2019.

[33] Natalia L Komarova. Replicator–mutator equation, universality property and population dynamics of learning. *Journal of Theoretical Biology*, 230(2):227–239, 2004.

[34] D.E. Koulouriotis and A. Xanthopoulos. Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Applied Mathematics and Computation*, 196(2):913 – 922, 2008.

[35] Kevin N Laland. Social learning strategies. *Animal Learning & Behavior*, 32(1):4–14, 2004.

[36] Jee Hang Lee, Ben Seymour, Joel Z. Leibo, Su Jin An, and Sang Wan Lee. Toward high-performance, memory-efficient, and fast reinforcement learning—lessons from decision neuroscience. *Science Robotics*, 4(26), 2019.

[37] Sang Wan Lee, Shinsuke Shimojo, and John P O'Doherty. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3):687–699, 2014.

[38] Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742*, 2019.

[39] M Lengyel and P Dayan. Hippocampal contributions to control: The third way. In *Twenty-First Annual Conference on Neural Information Processing Systems (NIPS 2007)*, pages 889–896. Curran, 2008.

[40] Jie Lin, Wei Yu, Nan Zhang, Xinyu Yang, Hanlin Zhang, and Wei Zhao. A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications. *IEEE Internet of Things Journal*, 4(5):1125–1142, 2017.

[41] TJH Morgan, LE Rendell, Micael Ehn, W Hoppitt, and Kevin N Laland. The evolutionary basis of human social learning. *Proceedings of the Royal Society B: Biological Sciences*, 279(1729):653–662, 2012.

[42] Wataru Nakahashi. The evolution of conformist transmission in social learning when the environment changes periodically. *Theoretical population biology*, 72(1):52–66, 2007.

[43] Peter Nemenyi. Distribution-free multiple comparisons. In *Biometrics*, volume 18, page 263. International Biometric Society, 1962.

[44] Nicolas Bredeche and Jean-Marc Montanier and Wenguo Liu and Alan FT Winfield. Environment-driven distributed evolutionary adaptation in a population of autonomous robotic agents. *Mathematical and Computer Modelling of Dynamical Systems*, 18(1):101–129, 2012.

[45] Martin A Nowak. *Evolutionary dynamics: exploring the equations of life*. Harvard university press, 2006.

[46] John O'Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J. Dolan. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669):452–454, 2004.

[47] Andreas Olsson, Ewelina Knapska, and Björn Lindström. The neural and computational systems of social learning. *Nature Reviews Neuroscience*, 21(4):197–212, 2020.

[48] John P O'Doherty, Sang Wan Lee, and Daniel McNamee. The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, 1:94–100, 2015.

[49] John P O'Doherty, Sangwan Lee, Reza Tadayonnejad, Jeff Cockburn, Kyo Iigaya, and Caroline J Charpentier. Why and how the brain weights contributions from a mixture of experts. *Neuroscience & Biobehavioral Reviews*, 2021.

[50] Luke Rendell, Robert Boyd, Daniel Cownden, Marquist Enquist, Kimmo Eriksson, Marc W Feldman, Laurel Fogarty, Stefano Ghirlanda, Timothy Lillicrap, and Kevin N Laland. Why copy others? insights from the social learning strategies tournament. *Science*, 328(5975):208–213, 2010.

[51] Michael Rubenstein, Alejandro Cornejo, and Radhika Nagpal. Programmable self-assembly in a thousand-robot swarm. *Science*, 345(6198):795–799, 2014.

[52] Debra Satz and John Ferejohn. Rational choice and social theory. *The Journal of philosophy*, 91(2):71–87, 1994.

[53] Karl H. Schlag. Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits. *Journal of Economic Theory*, 78(1):130 – 156, 1998.

[54] John Scott. Rational choice theory. *Understanding contemporary society: Theories of the present*, 129:671–85, 2000.

[55] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge university press, 1982.

[56] Kenneth O Stanley, Jeff Clune, Joel Lehman, and Risto Miikkulainen. Designing neural networks through neuroevolution. *Nature Machine Intelligence*, 1(1):24–35, 2019.

[57] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[58] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[59] Michel Tokic. Adaptive ε-greedy exploration in reinforcement learning based on value differences. In *Annual Conference on Artificial Intelligence*, pages 203–210. Springer, 2010.

[60] Wataru Toyokawa, Hye-rin Kim, and Tatsuya Kameda. Human collective intelligence under dual exploration-exploitation dilemmas. *PloS one*, 9(4):e95789, 2014.

[61] Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860–868, 2018.

[62] Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N Ahmad Aziz, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature*, pages 1–7, 2021.

[63] Mike M Webster and KN Laland. Social learning strategies and predation risk: minnows copy only when using private information would be costly. *Proceedings of the Royal Society B: Biological Sciences*, 275(1653):2869–2876, 2008.

[64] Andrew Whiten. The burgeoning reach of animal culture. *Science*, 372(6537), 2021.

[65] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[66] Anil Yaman and Giovanni Iacca. Distributed embodied evolution over networks. *Applied Soft Computing*, 101:106993, 2021.

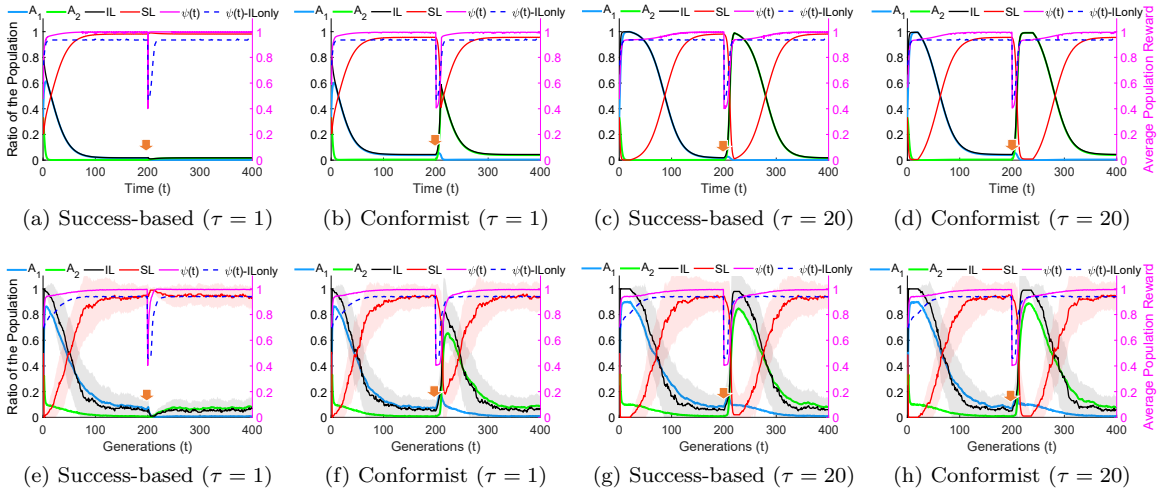
[67] Anil Yaman, Decebal Constantin Mocanu, Giovanni Iacca, George Fletcher, and Mykola Pechenizkiy. Limited evaluation cooperative co-evolutionary differential evolution for large-scale neuroevolution. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 569–576, 2018.

A Supplementary Material

A.1 Population dynamics

We present the results of the mathematical model at the top row of Figure 8 and the evolutionary algorithm at the bottom row of Figure 8. Notably, these approaches produce very similar results. The social learners using the success-based strategy show dominance over individual learners throughout the learning process as well as rapid adaptation after the reward reversal. On the other hand, the social learners using the conformist strategy show dominance only after the majority of the individuals learn to make optimum choices. When the latency is increased, the success-based strategy shows similarity to the results of the conformist strategy.

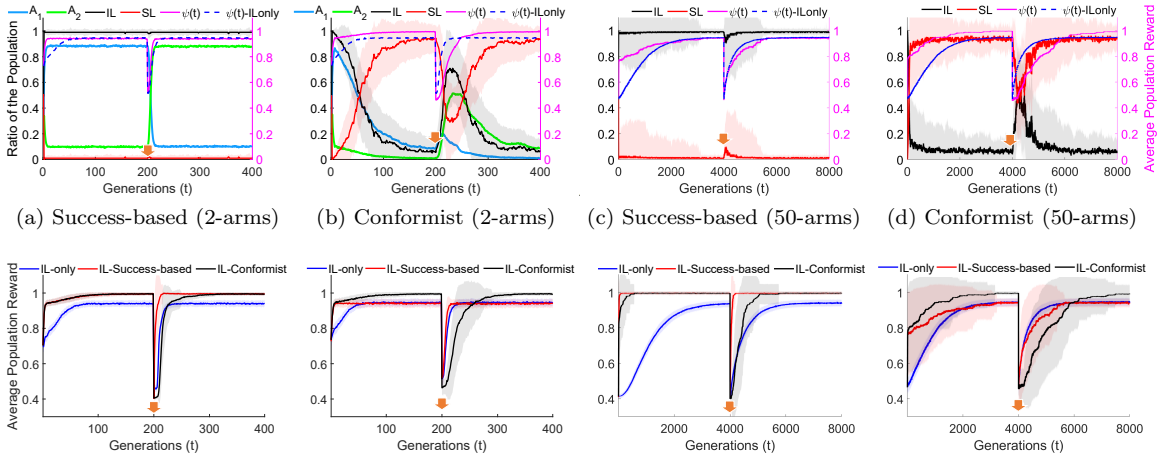
In highly uncertain environment, the success-based social learners lose their dominance in the population throughout the evolutionary processes (see Figure 9). This indicates that the fitness of individual learners is higher than that of the success-based social learners. On the other hand, the behavior of conformist learners in uncertain environments is similar to that in static environments.



(a) Success-based ($\tau = 1$) (b) Conformist ($\tau = 1$) (c) Success-based ($\tau = 20$) (d) Conformist ($\tau = 20$)

(e) Success-based ($\tau = 1$) (f) Conformist ($\tau = 1$) (g) Success-based ($\tau = 20$) (h) Conformist ($\tau = 20$)

Figure 8: The population dynamics produced by the mathematical model ((a) through (d), see Section 4.5.1) and evolutionary algorithm ((e) through (h), see Section 4.5.2) on a 2-armed bandit (binary decision-making) task with reward distributions $R_1 \sim \mathcal{N}(1, 0.05)$ and $R_2 \sim \mathcal{N}(0.4, 0.05)$. Figures in the first two columns show the results when the latency (τ) for the social learners equals to 1, whereas, the figures in last two column show the results when it is set to 20. The x and left y axes show the ratio of individual and social learners in the population, and the right y shows the average population reward (fitness) $\psi(t)$. $A_1(t)$ and $A_2(t)$ show the ratios of the individual learners that chose the first and second arms at t respectively ($A_1(t) + A_2(t) = IL(t)$).



(a) Success-based (2-arms) (b) Conformist (2-arms) (c) Success-based (50-arms) (d) Conformist (50-arms)

(e) Low uncertainty (2-arms) (f) High uncertainty (2-arms) (g) Low uncertainty (50-arms) (h) High uncertainty (50-arms)

Figure 9: Figures (a) through (d) show the population dynamics of the evolutionary processes in uncertain environments with 2 and 50 arms (visualization of the ratios of the individual arms in 50 armed case is omitted due to the large number of arms). The uncertainty is introduced by increasing the standard deviation of the reward distribution of sub-optimum arm (optimum distribution: $R_1 \sim \mathcal{N}(1, 0.05)$, and sub-optimum distribution: $R_2 \sim \mathcal{N}(0.4, 0.5)$).
Figures (e) through (h) show the average population reward ($\psi(t)$) in populations consisting of only individual learners, individual learners with success-based social learners, and individual learners with conformist social learners in environments with low and high uncertainty with 2 and 50 arms. Highlighted areas indicate the standard deviations of independent runs of the evolutionary algorithm. In all figures, orange arrows mark the reward reversal where the optimum and sub-optimum reward distributions are swapped.

A.2 Training environment

Figure 10 illustrates the environment used for training phase of the algorithms. The trained algorithms then tested on separate environments and reported in the main text of the paper.

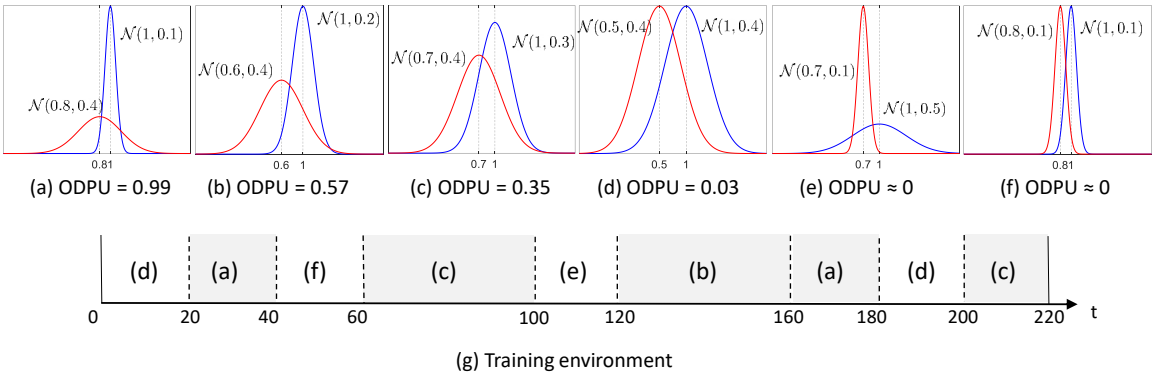


Figure 10: The environment used for training processes for algorithms: SL-GA and SL-NE. (a) through (f) show arbitrarily defined reward distributions and their ODPUs for 2 arms. (g) shows the training environment generated by using the specified reward distributions for specified lengths of periods. The complete period consists of 220 time steps. Dashed vertical lines indicate the change of the reward distribution points. The periods with uncertainty (ODPU$\geq$ 0.1) are highlighted in gray.

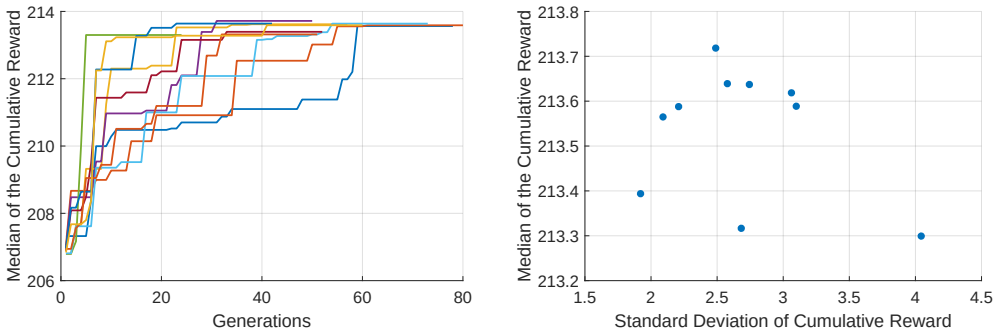
A.3 Evolutionary optimization of the SL-GA

In this section, we provide the results of the evolutionary processes of the GA used to optimize the decision policies in the SL-GA. We performed 10 independent GA runs on the environment shown in Figure 10. Fitness values are the median of the cumulative rewards of 112 processes.

The fitness trends during the GA processes, and the fitness versus standard deviations of the best solution (SL-GA policies) are shown in Figure 11.

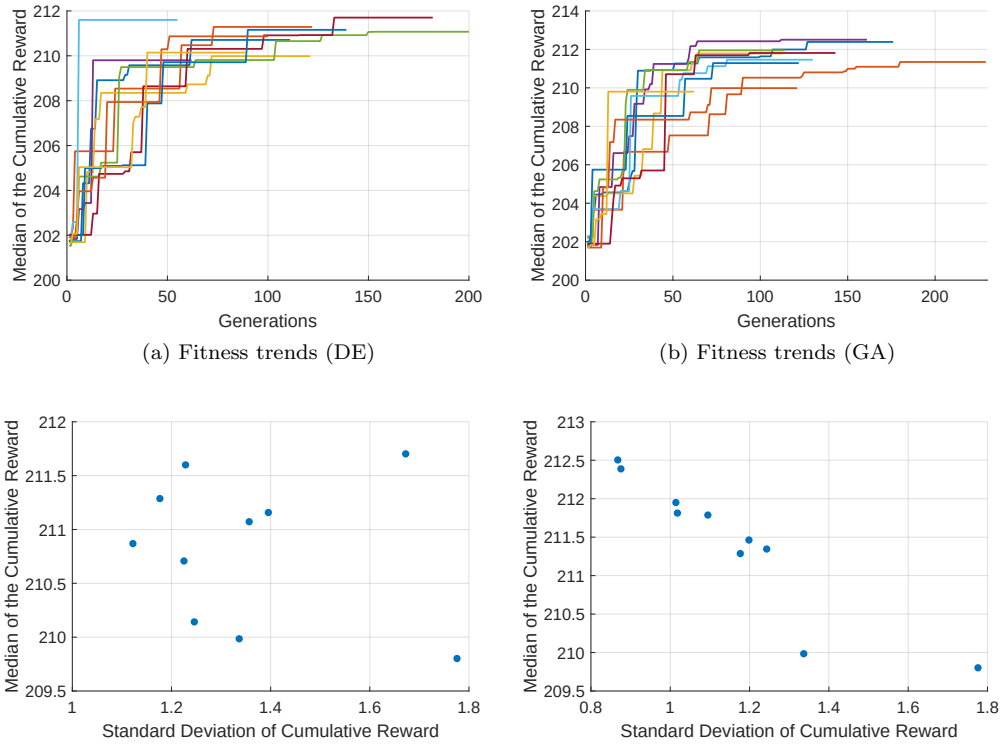
A.4 Evolutionary optimization of the SL-NE

Figure reffig:runNE shows the evolutionary optimization processes of the ANN based policies using genetic algorithms and differential evolution. The evolutionary processes are terminated if there is no fitness improvement for 50 consecutive generations.



(a) Fitness trends (b) Fitness vs. standard deviation

Figure 11: (a) Fitness trend of 10 independent GA runs, and (b) fitness versus standard deviations of the best solution for each independent GA runs.



(a) Fitness trends (DE) (b) Fitness trends (GA)

(c) Fitness vs. standard deviation (DE) (d) Fitness vs. standard deviation (GA)

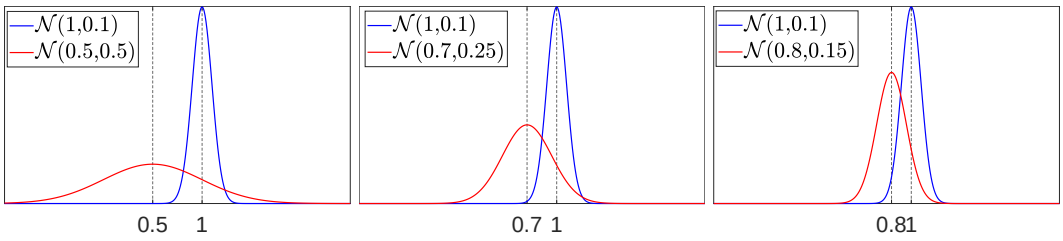
Figure 12: The evolutionary process of the ANNs using Neuroevolution approach. (a) and (b) fitness trends of 10 independent DE and GA runs, and (c) and (d) fitness versus standard deviations of the best solution for each independent DE and GA runs.

A.5 Experiment design and results

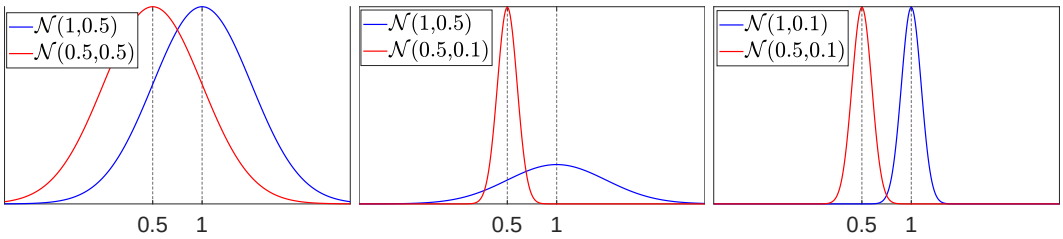
We designed three sets of experiments with various volatility and uncertainty in terms of environment change and the overlap between the reward distributions. In **Experiment1**, we designed four environments as: stable low uncertainty, stable high uncertainty, volatile low uncertainty and volatile high uncertainty. In stable environments, the reward distributions were changed two, and in volatile environments five times. We defined six reward distributions (shown in Figure 13) and assigned to a time period in the processes as illustrated in Figures 15a, 15b, 15c and 15g. Low and high uncertainty environments have low and high ODPUs respectively.

In **Experiment2** (random volatile environment), we defined random environments by selecting the number of environment change between $[10, 30]$ from uniform distribution, and assigned a reward distribution (from Figure 13) to each period between environment change points randomly.

In **Experiment3** (gradual environment), we defined gradual environment change by defining the reward distributions based on sinusoidal functions as shown in Figure 14.



(a) ODPU = 0.97 (b) ODPU = 0.59 (c) ODPU = 0.14



(d) ODPU = 0.10 (e) ODPU $\simeq$ 0 (f) ODPU $\simeq$ 0

Figure 13: The reward distributions of the arms defined in each period of the environments shown in Figures (a), (b), (c) and (g).

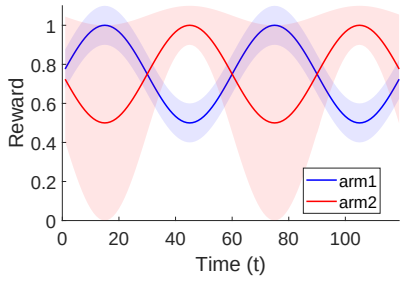


Figure 14: Reward distributions and their standard deviations (highlighted) in gradual environment. We used two sinusoidal functions to model environment change. The standard deviation of the first arm kept constant while an additional sinusoidal function is used to model the change in the standard deviations of the second arm depending on time.

Figure 15 show the average and cumulative average population rewards obtained by the meta-social learning strategies in all experiments. Figures that show the performance versus exploration cost, and ranks of the strategies are given in Figure 15.

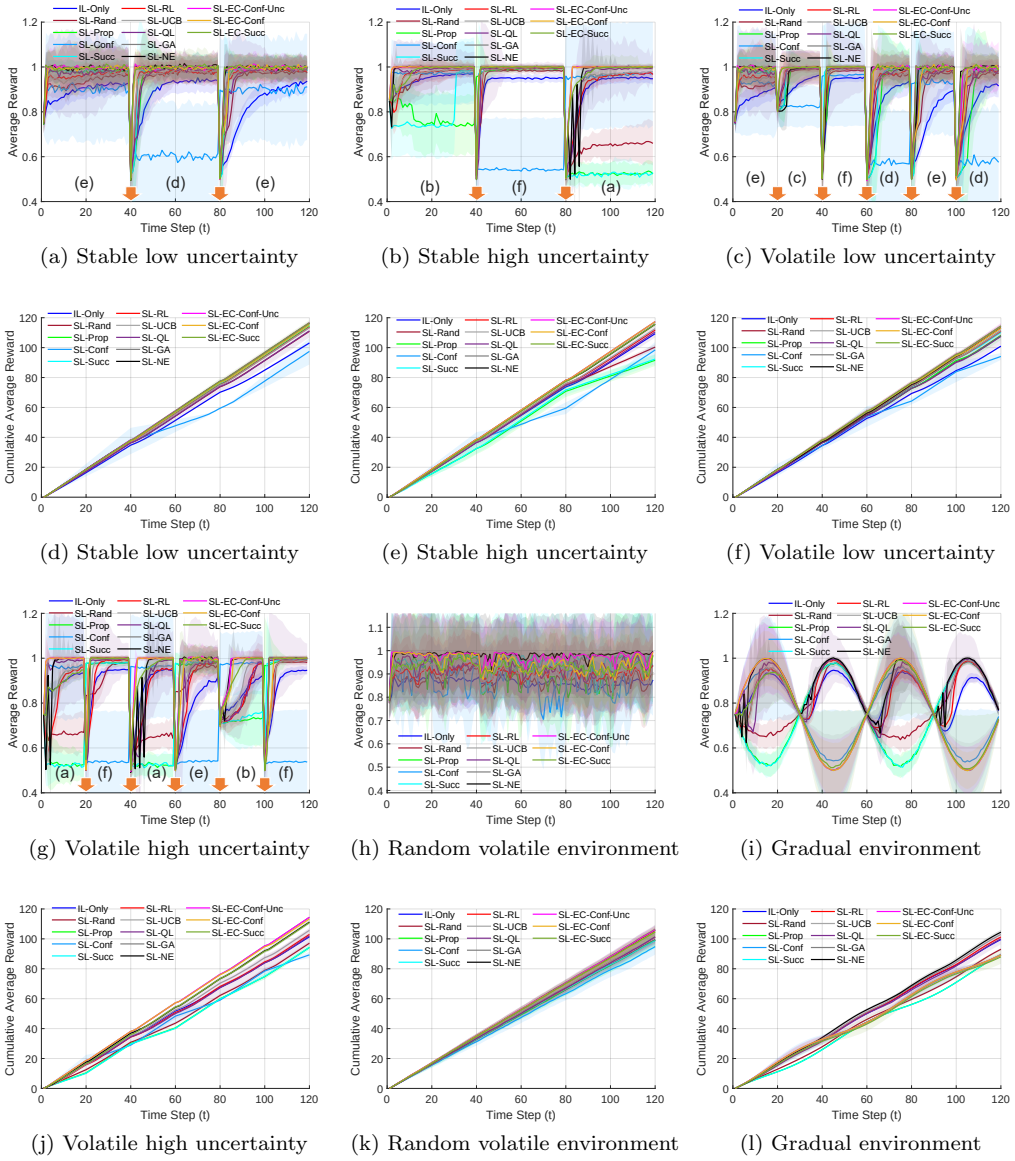
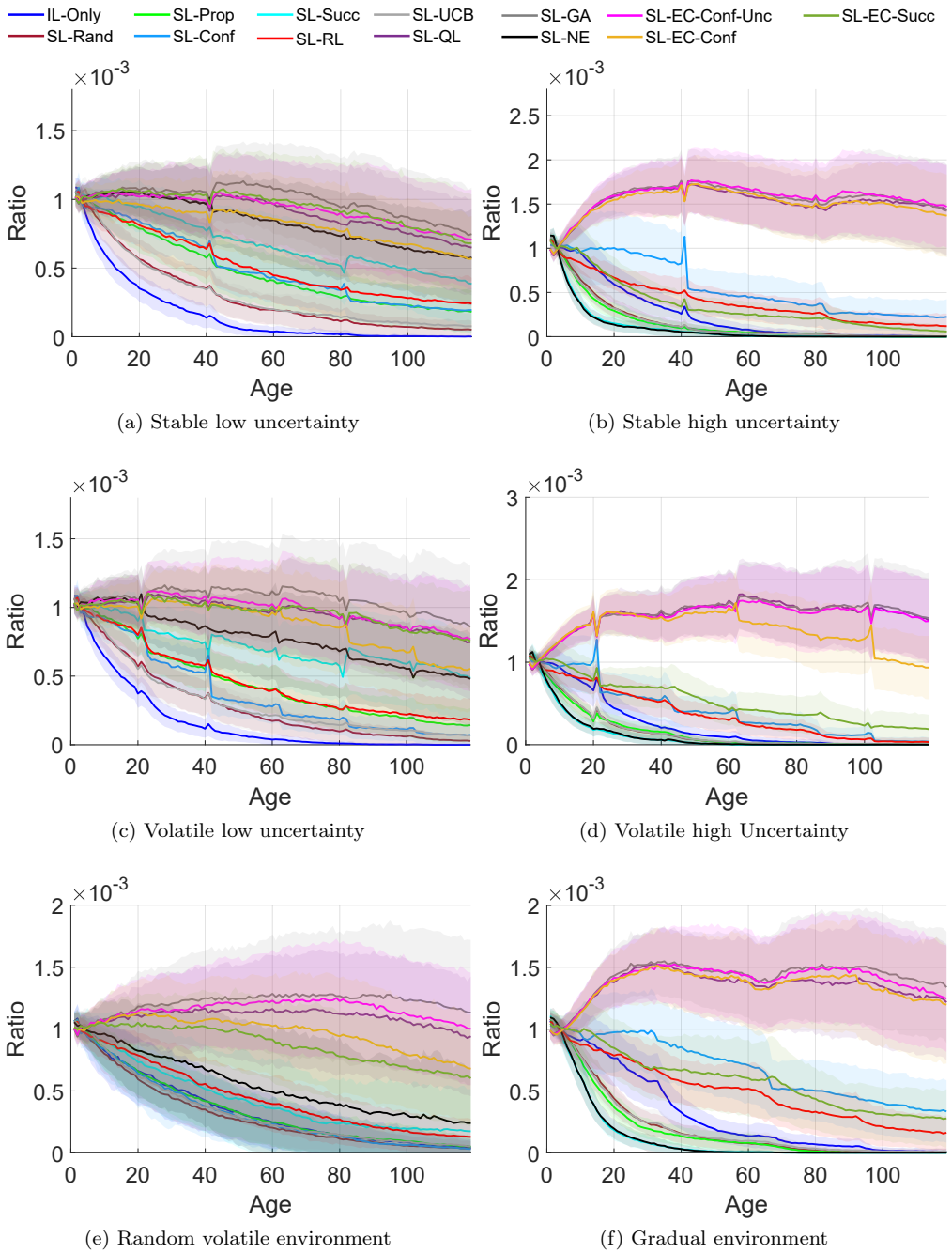


Figure 15: The average and cumulative average population rewards obtained by the meta-social learning strategies throughout the processes.
The average results and standard deviations (highlighted) are based on 112 independent runs of each algorithm. Each orange arrow in x-axes indicates a reward reversal time point.

A.6 Evolution of meta-social learners: age distributions



(a) Stable low uncertainty

(b) Stable high uncertainty

(c) Volatile low uncertainty

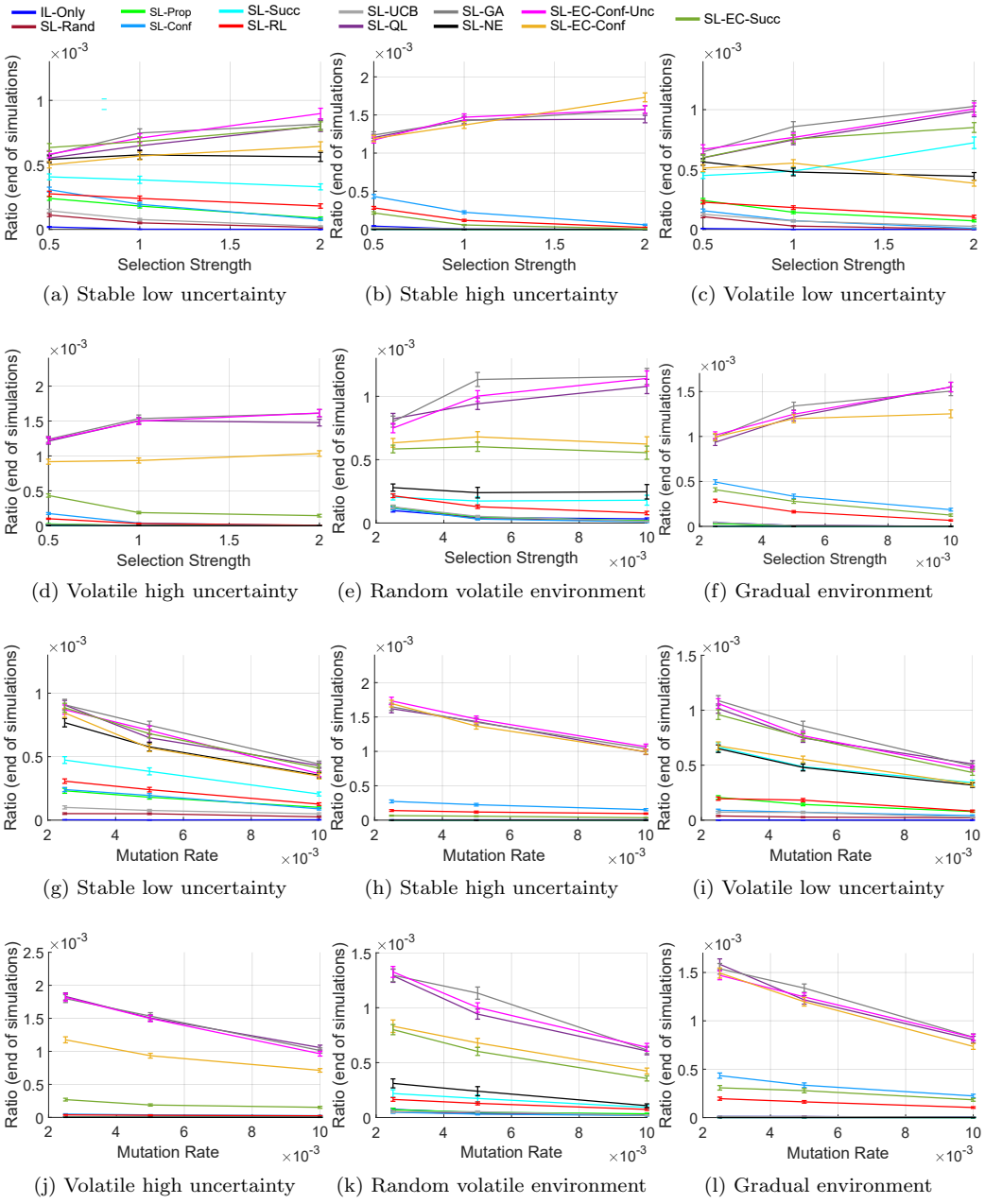
(d) Volatile high Uncertainty

(e) Random volatile environment

(f) Gradual environment

Figure 16: The age distributions of the meta-social strategies throughout the evolutionary processes. The age distributions of the dominant strategies are higher relative to others.

A.7 Evolution of meta-social learners: sensitivity analysis



(a) Stable low uncertainty (b) Stable high uncertainty (c) Volatile low uncertainty

(d) Volatile high uncertainty (e) Random volatile environment (f) Gradual environment

(g) Stable low uncertainty (h) Stable high uncertainty (i) Volatile low uncertainty

(j) Volatile high uncertainty (k) Random volatile environment (l) Gradual environment

Figure 17: The effect of selection strength and mutation rate to the domination of the successful strategies. While the selection strength increases, domination of successful strategies increases; however, while the mutation rate increases, domination of successful strategies decreases.