CONFORMITY LEADING SELF-ORGANISED RESCUE FROM A SUBOPTIMAL DECISION-MAKING BIAS

A PREPRINT

Wataru Toyokawa*
Department of Psychology,
University of Konstanz, Germany

Wolfgang Gaissmaier
Department of Psychology,
Centre for the Advanced Study of Collective Behaviour,
University of Konstanz, Germany

March 19, 2021

ABSTRACT

Given the ubiquity of potentially suboptimal biases observed in animal behaviour, it seems paradoxical that improvements of decision-making performance through the conformist-based social learning, a process widely considered as bias-amplification, still prevails. Here we show, through model analyses and online experiments with 467 adult human subjects, that conformity can promote adaptive risk-taking in repeated decision-making, even though many individuals are potentially biased toward suboptimal risk-aversion due to myopia of reinforcement learning. Although positive feedback imposed by conformity could result in maladaptive informational cascades, our behavioural dynamical model identified a key role of negative feedback that arises when a weak minority influence undermines the inherent behavioural bias. This 'collective behavioural rescue', emerging through coordination of positive and negative feedback, highlighted a benefit of social learning in wider environmental conditions than previously assumed, and resolved the ostensible paradox of flexibility through conformity under risk and uncertainty.

Keywords risk aversion · social learning · collective behavioural rescue · the hot stove effect · self-organisation · decisions from experience · conformity · collective behaviour

*Corresponding author: wataru.toyokawa@uni-konstanz.de

1 Introduction

Although the potential cost of strong social influences in decision-making has been well recognised (Boyd and Richerson, 1985; Kandler and Laland, 2013; Kendal et al., 2005; King and Cowlishaw, 2007; Raafat et al., 2009; Giraldeau et al., 2002; Toyokawa et al., 2019), collective learning with the conformist-biased copying tendency has been considered widely adaptive in a range of environmental conditions (Rendell et al., 2010; Kameda and Nakanishi, 2002; Toyokawa et al., 2014, 2019; Miu et al., 2018). One rationale behind this optimistic view might come from an assumption that individuals should prefer a behavioural option that brings larger net benefits in utility to those bringing lower outcomes. Therefore, even though individual decision-making is fallible due to uncertainty, this should be such a noise that the statistical filtering through informational pooling can cancel out by the low of large numbers. As a consequence, the group as a whole can boost its probability to choose the optimal behaviour.

However, as numerous examples in behavioural science have documented, animals do not always tend to choose or prefer the benefit-maximising behaviour over other behavioural alternatives. Instead, various behavioural biases, such as risk-aversion or uncertainty-avoidance, have been identified in many different decision-making contexts (Hertwig and Erev, 2009; Real and Caraco, 1986; Frey et al., 2017; Yechiam et al., 2006; Weber, 2006), which may induce a suboptimal decision-making tendency in a particular context. Therefore, the ostensible prerequisite of collective intelligence, that is, individuals must possess a higher chance to exhibit a better behaviour than choosing a suboptimal alternative, may not always hold in reality. What is worse, social influences may nourish the spread of those suboptimal biases (Chung et al., 2015; Dussutour et al., 2005; Suzuki et al., 2016; Shupp and Williams, 2008; Jouini et al., 2013; Moussaïd et al., 2015).

Given that the biases are ubiquitous and individual animals hardly escape from them, it may seem that collective intelligence through social influences cannot be realised in many actual decision-making situations where individuals exhibit suboptimal biases. However, the collective improvement of decision accuracy have been widely observed in the real world in a wide range of taxa (Sasaki and Biro, 2017; Seeley et al., 1991; Alem et al., 2016; Sumpter, 2006; Camazine et al., 2001), including very animals known to exhibit suboptimal risk-taking biases such as bumblebees (Real et al., 1982), honeybees (Drezner-Levy and Shafir, 2007), and pigeons (Ludvig et al., 2014). How, if any, can group-living animals improve collective decision accuracy while suppressing the potentially deleterious effect of fundamental decision-making biases?

Here we provide a mathematical model and experimental results of a collective learning and decision-making task. We assumed that decision-makers could repeatedly choose between a safe alternative that gave certain payoffs and a risky alternative that gave uncertain payoffs, while they could observe other players' choices (McElreath et al., 2005, 2008; Toyokawa et al., 2014, 2017, 2019; Deffner et al., 2020). To maximise long-term individual profit of their own in such an experience-based risky decision-making task, it is crucial to strike a right balance between exploiting the option that has seemed better so far versus exploring the other options to seek informational gain. Due to the nature of adaptive information sampling under such exploration-exploitation trade-offs, lone decision-makers often end up in risk-aversion, trying to reduce the chance of further failures once the individual has experienced an unfavourable outcome from the risky alternative (March, 1996; Denrell, 2007; Hertwig and Erev, 2009), a phenomenon known as the *hot stove effect*. Risk-aversion is suboptimal in the long run if the risky arm gives higher payoffs on average (Denrell and March, 2001).

The hot stove effect can be mitigated if individuals can learn not only from payoffs of the chosen option but also from forgone outcomes of the unchosen alternative (Denrell, 2007; Yechiam and Busemeyer, 2006). Therefore, if payoff information of other individuals is made publicly observable among the group, it would not be surprising that learning from the public information rescues individuals from suboptimal risk-aversion because the public information could act as the forgone payoffs. However, such complete public information may not be readily available, and only behavioural cues, such as frequency of choices indicating how many other individuals choose each behavioural alternative, may be accessible in many realistic ecological circumstances. Therefore, if collective intelligence in risky decision-making can widely emerge even under the potential danger of suboptimal hot stove effect, it should be a simple, behavioural-cue-based mechanism that alters the course of learning and mitigates the suboptimal bias.

Using an agent-based computational model simulation and analyses of a differential equation model representing approximated dynamics of collective behaviour, here we showed theoretically that such the *collective behavioural rescue* could indeed emerge through self-organisation driven by the frequency-based social influences. The theoretical predictions were further supported by online behavioural experiments with 467 adult human subjects. We argue that mutual influences among social learners can alter, rather than amplify, the suboptimal behavioural bias, suggesting a previously overlooked benefit of conformity.

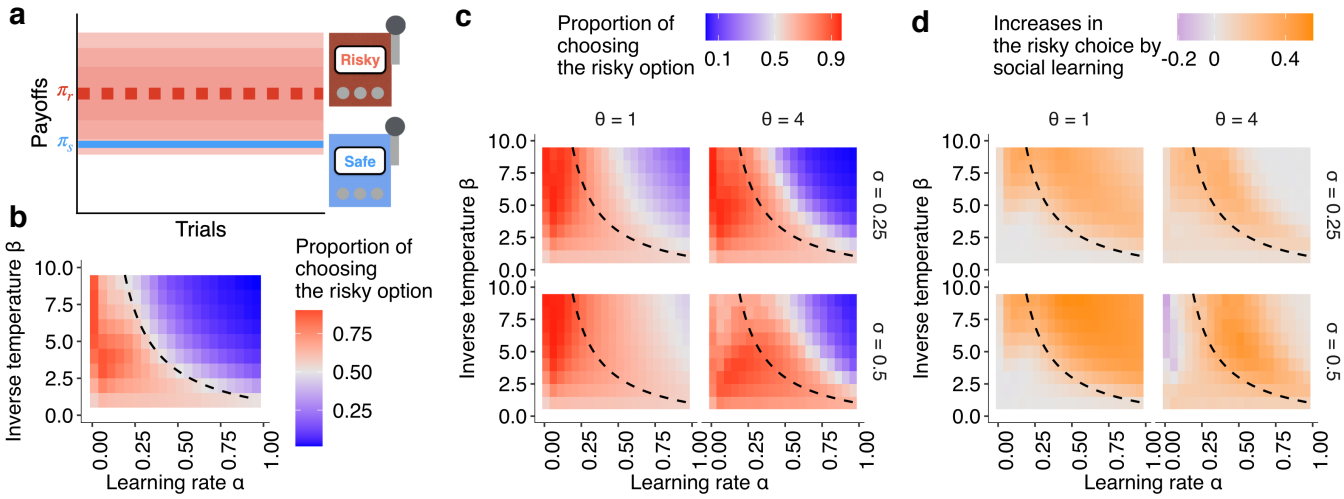

Figure 1: Mitigation of the hot stove effect by collective decision-making in the two-armed risky bandit task. (a) A schematic diagram of the task. A safe option gives a constant reward $\pi_s = 1$ whereas a risky option gives a reward randomly drawn from a Gaussian distribution with mean $\mu = 1.5$ and s.d. $= 1$. (b, c): The emergence of risk-taking bias due to the hot stove effect depending upon a combination of the reinforcement learning parameters; (b): under no social influence (i.e. the copying weight $\sigma = 0$), and (c): under social influences with different degrees of conformity exponents θ and copying weights σ. The dashed curve is the asymptotic equilibrium at which asocial learners are expected to end up choosing both alternatives equally likely (i.e. $P_r^\star = 0.5$) which is given analytically by $\beta = (2 - \alpha)/\alpha$. The coloured background is a result of the agent-based simulation with total trials $T = 150$, showing the average proportion of choosing the risky option in the second half of the learning trials $P_{r,t>75} > 0.5$ under a given combination of the parameters. (d): The differences between the mean proportion of risk-aversion of asocial learners and that of social learners, highlighting regions in which performance is improved (orange) or undermined (purple) by social learning.

2 Results

2.1 The collective behavioural rescue effect

The collectively-induced mitigation of suboptimal risk-aversion was observed in the simplest two-armed bandit task where a formal analysis had been available (Denrell, 2007). Consider a decision-maker facing a repeated choice between a safe (i.e. certain) alternative s and a risky (i.e. uncertain) alternative r for over T trials (Fig. 1a). We followed Denrell (2007)'s setup equipped with a safe option generating a constant payoff (which was 1 in Fig. 1), coupled with a risky option whose payoff π_r followed a Gaussian distribution with mean $= \mu$ and s.d. $= s$ (Fig. 1). In the main anal-

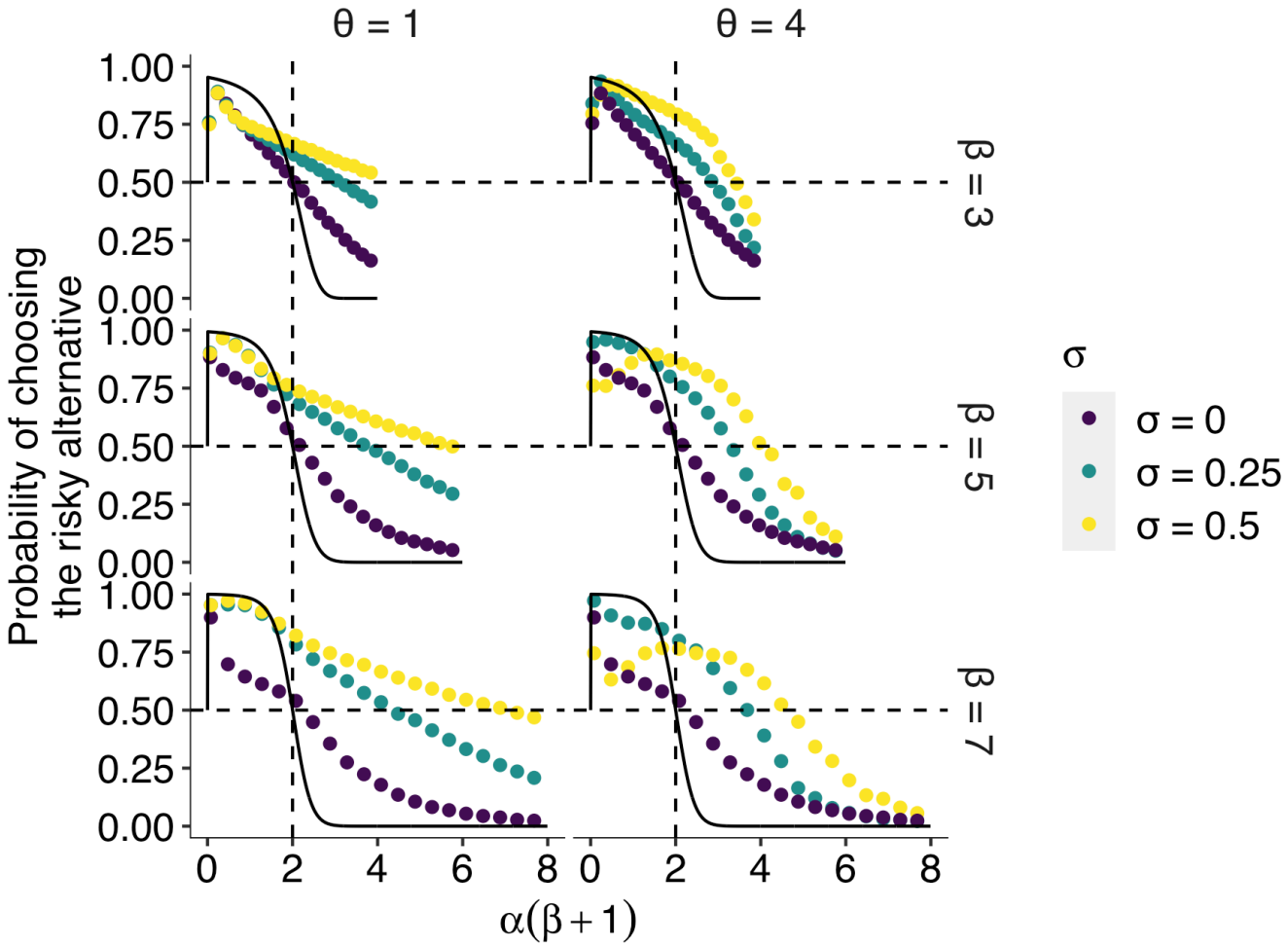

Figure 2: The effect of social learning on the average decision performance. The x-axis is a product of two reinforcement learning parameters $\alpha(\beta+1)$, namely, the susceptibility to the hot stove effect. The y-axis is the mean proportion of choosing the optimal risky alternative in the last 75 trials in the two-armed bandit task whose setup was the same as in Fig. 1. The black solid curve is the analytical prediction of the asymptotic performance of individual reinforcement learning with infinite time horizon $T \to +\infty$. The analytical curve shows a choice shift emerging at $\alpha(\beta+1) = 2$, that is, individual learners ultimately prefer the safe- to the risky-option when $\alpha(\beta+1) > 2$ holds in the current setup of the task. The dotted curves are mean results of agent-based simulations of social learners with two different mean values of the copying weight $\sigma \in \{0.25, 0.5\}$ (green and yellow, respectively) and asocial learners with $\sigma = 0$ (purple). The difference between the agent-based simulation with $\sigma = 0$ and the analytical result was due to the finite number of decision trials in the simulation, and hence, the longer the horizon, the closer they become (Supplementary Fig S 8). Each panel shows different combinations of the inverse temperature β and the conformity exponent θ.

ysis, we focused on the case where the risky alternative had $\mu = 1.5$ and $s = 1$, that is, choosing the risky alternative was the optimal strategy for a decision-maker in a long run.

In such a risk-heterogenous multi-armed bandit setting, reinforcement learners are prone to exhibit suboptimal risk-aversion (March, 1996; Denrell, 2007; Hertwig and Erev, 2009), even though they could have achieved high performance in a risk-homogenous task where all options have an equivalent payoff variance (Sutton and Barto, 2018). Denrell (2007) mathematically derived a condition under which the suboptimal risk-aversion arises, depicted by the dashed curve in Fig. 1b. The suboptimal risk-aversion becomes prominent when the value-updating in learning is myopic (that is, the learning rate α is large) and/or action-selection is exploitative (that is, the inverse temperature of the softmax action selection β is large). Under such circumstances, experiences of low-value payoffs from the risky option tend to repel the decision-maker from further choosing the risky option, making her trapped in the safe alternative, a phenomenon known as the hot stove effect. In sum, whenever the interaction between the two learning parameters $\alpha(\beta + 1)$ exceeds a threshold value, which was 2 in the current example, decision-makers are expected to become aversive to the risky option (Fig. 2).

As shown in Fig. 1c, a simple frequency-based social influence, even with a conformist-biased influence ($\theta > 1$), widened the region of parameter combinations where decision-makers can escape from the suboptimal risk aversion (the red area). The increment of the area of adaptive risk-seeking was wider with $\theta = 1$ than $\theta = 4$. When $\theta = 1$, a large copying weight σ can eliminate almost all the area of risk-aversion (Fig. 1c; see also Supplementary Fig.7 for wider parameter combinations). Note that increasing copying weight $\sigma \to 1$ eventually approximates choices to a random decision-making (Supplementary Fig.7). Interestingly, such an alternation to risk-seeking did not emerge when risk-aversion was actually optimal (Supporting Figure 11), suggesting that social influence does not mindlessly increase risk-seeking; instead, it may work only when to do so is adaptive.

Fig. 2 highlights how widely the risk-aversion was relaxed through social influences. Individuals with positive $\sigma > 0$ could maintain the high proportion of risk-seeking even in the region of susceptible to the hot stove effect $\alpha(\beta+1) > 2$. Although social learners eventually fell into the risk-aversion regime with increasing $\alpha(\beta+1)$, risk-aversion was widely mitigated compared to the performance of individual learners who had $\sigma = 0$. Interestingly, the probability of choosing the optimal risky option was maximised at an intermediate value of $\alpha(\beta+1)$ when conformity exponent was large $\theta = 4$ and copying weight was high $\sigma = 0.5$.

In the region of less susceptible to the hot stove effect ($\alpha(\beta+1) < 2$), social influence could enhance individual optimal risk-seeking up to the theoretical benchmark expected in sole reinforcement learning with an infinite time horizon. The

socially-induced increment of risk-seeking in the region $\alpha(\beta + 1) < 2$ was more evident with larger β, and hence with smaller α in order to to satisfy $\alpha(\beta + 1) < 2$. The smaller the learning rate α, the longer it would take to achieve the asymptotic equilibrium state due to the slow value-updating. Asocial learners, as well as social learners with high σ (= 0.5) coupled with high θ (= 4), were still far from the analytical benchmark, whereas social learners with weak social influence $\sigma = 0.25$ were able to converge nearly on the benchmark performance, suggesting that social learning might affect the speed of learning. Indeed, a longer time horizon $T = 1075$ reduced the advantage of weak social learners in this $\alpha(\beta + 1) < 2$ region because slow learners could now achieve the benchmark accuracy (Supplementary Fig. 8; Fig. 9).

The approaching to the benchmark by the elongated time horizon, and the concomitant reduction of advantage of social learners, was also found in the high susceptible region $\alpha(\beta + 1) \gg 2$ especially for those who had a high conformity exponent $\theta = 4$ (Supplementary Fig. 8). Notably, however, the facilitation of optimal risk-seeking became further evident in the other intermediate region $2 < \alpha(\beta + 1) < 4$. This suggests that merely speeding-up or slowing-down of learning could not satisfactorily account for the qualitative choice-shift emerging through social influences.

We obtained similar results across different settings of the multi-armed bandit tasks such as skewed payoff distribution from which large payoffs realised less likely whereas small payoffs realised commonly due to an asymmetric probability unlike the Gaussian noise (March, 1996; Denrell, 2007) (Supporting Figure S12). Further, the conclusion still held under different assumptions of social influences on reinforcement learning, assuming that the conformity-biased influence acts on the learning process (the value-shaping model (Najar et al., 2020)) rather than on the action-selection (the decision-biasing model) assumed above (Supporting Methods). Although groups of agents with the value-shaping algorithm seemed more prone to herding, the results were not qualitatively changed and collective behavioural rescue emerged when β is sufficiently small (Supporting Figure S10).

2.2 The effect of individual heterogeneity

We have thus far assumed little parameter variations across individuals in a group to focus on the qualitative differences between social and asocial learners' behaviour. However, individual differences in development, state, experience or variation in behaviour caused by personality traits might either facilitate or undermine collective decision performance. Especially, considering a group composed of both types of individuals who are less susceptible to the hot stove effect $\alpha(\beta + 1) < 2$ as well as those who are more susceptible $\alpha(\beta + 1) > 2$, it remains unclear who receives the benefit of

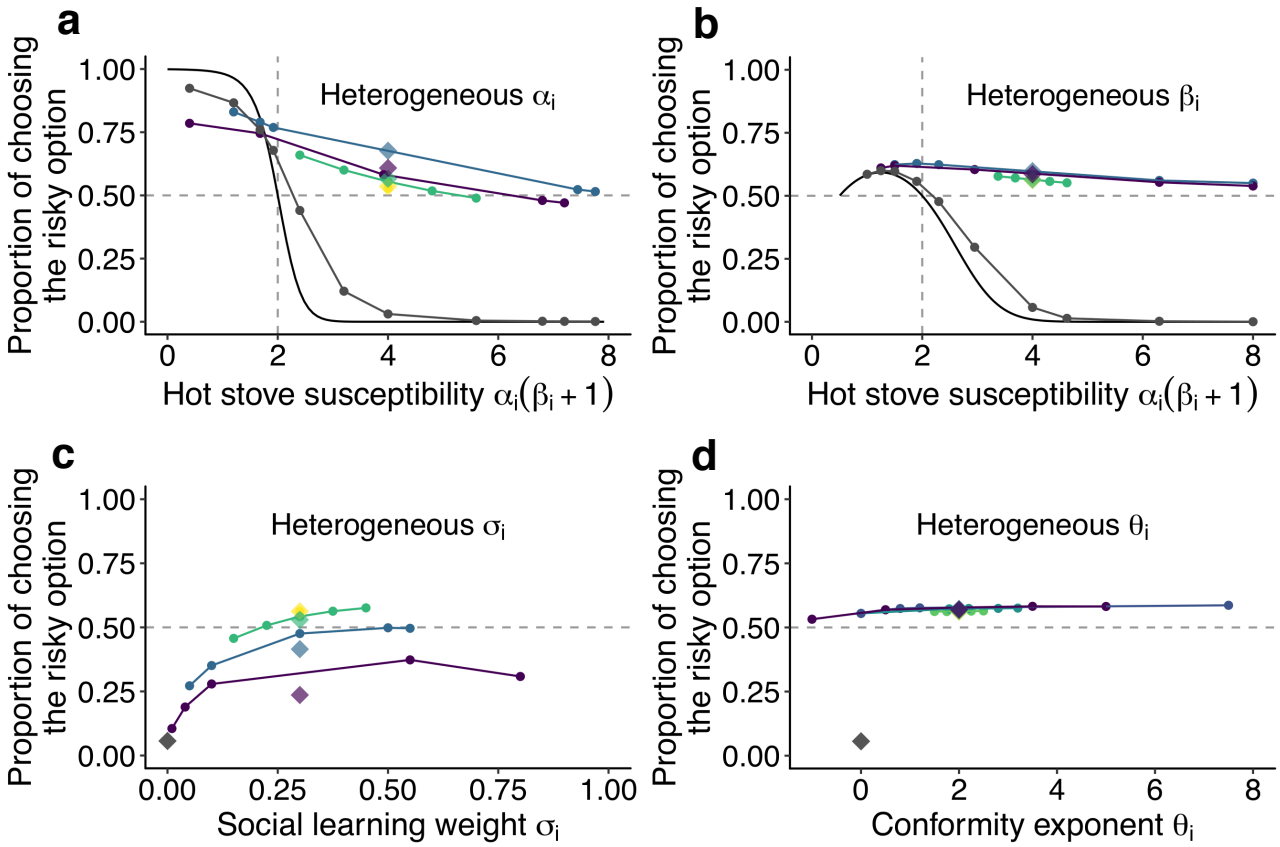

Figure 3: The effect of individual heterogeneity on the proportion of choosing the risky option in the two-armed bandit task. (a) The effect of heterogeneity of α, (b) that of β, (c) that of σ, and (d) that of θ. Individual values of a focal behavioural parameter were varied across individuals in a group of five. Other non-focal parameters were identical across individuals within a group. The basic parameter values assigned to non-focal parameters were $\alpha = 0.5$, $\beta = 7$, $\sigma = 0.3$, and $\theta = 2$, and groups' mean values of the focal varying parameters were matched to these basic values. We simulated 3 different heterogeneous compositions: The majority (3 out of 5 individuals) potentially suffer the hot stove effect $\alpha_i(\beta_i + 1) > 2$ (a, b) or the highest diversity for social learning parameters (c, d; purple); The majority could overcome the hot stove effect $\alpha_i(\beta_i + 1) < 2$ (a,b) or moderate heterogeneity for the social learning parameters (c,d; blue); and all individuals have $\alpha_i(\beta_i + 1) > 2$ but the smaller heterogeneity (green). The yellow shows the homogeneous groups' performance. Lines are drawn within average results across the same compositional groups. Each round-shaped dot represents each group member's mean performance. The diamonds are the average performance of each group composition. For a comparison, asocial learners' performance, from which the performance of social learners can be evaluated, was shown in grey. For the heterogeneous α and β, the analytical solution of asocial learning performance is shown as a solid-line curve. We ran 20,000 replications for each different group composition.

the rescue effect: Is it only those individuals with $\alpha(\beta + 1) > 2$ who enjoy the benefit, or can collective intelligence benefit a group as a whole?

Fig. 3a shows the effect of the heterogeneity in the learning rate α. Heterogeneous groups (that are shown either in green, blue, or purple) performed better on average than a homogeneous group (shown in the yellow diamond). The heterogeneous groups owe this overall improvement to the large rescue effect operating for highly susceptible individuals who have $\alpha(\beta + 1) \gg 2$. On the other hand, the performance of less susceptible individuals ($\alpha(\beta + 1) < 2$) were slightly undermined compared to the asocial benchmark performance. Notably, however, how large the detrimental effect for the small susceptible individuals depends upon the group's composition: The undermining effect was largely mitigated when low-susceptible individuals ($\alpha(\beta + 1) < 2$) composed a majority of a group (3 out of 5; the blue line); while they performed worse than the asocial benchmark when the majority was those with high susceptibility (purple).

The similar pattern was found for the inverse temperature β, whereas the impact of its heterogeneity was much smaller than that in α, and there was little difference between homogeneous and heterogeneous groups (Fig. 3b). Interestingly, no detrimental effect for less-susceptible individuals was found in association with the β variations.

On the other hand, individual variations in the copying weight σ had an overall detrimental effect on collective performance, although individuals in the highest diversity group could still perform better than the asocial benchmark (Fig. 3c). Individuals who had an intermediate level of copying weight achieved relatively higher performance within the group than those who had either higher or lower weight. This was because individuals with lower σ could lesser benefit from social information, while those with higher σ relied so heavily on social frequency information that behaviour was barely informed by individual learning. As a result, the average performance decreased with increasing diversity in σ.

Such a substantial effect of individual differences was not observed in the conformity exponent θ (Fig. 3d), where individual performance was almost stable regardless of whether the individual was a heavily conformist ($\theta_i = 8$) or even anti-conformist ($\theta_i = -1$). An existence of a few conformists in a group could not itself trigger a positive feedback among the group unless other individuals also relied on social information in the conformist-biased way, because flexible behaviour of non-conformists could keep the group's distribution nearly flat (that is, $N_s \approx N_r$). Therefore, the existence of individuals with small θ in a heterogeneous group could prevent the strong positive feedback from being immediately elicited, compensating the potential detrimental effect of maladaptive herding by strong conformists.

In sum, the relaxation of, and possibly the complete rescue from, the suboptimal bias in repeated risky decision-making widely emerged in a range of conditions in collective learning. It was not likely a mere speeding-up or slowing-down of learning process (Supplementary Fig. 9), nor just an averaging process mixing performances of both risk-seekers and risk-aversive individuals (Fig. 3). It did not depend on a specific characteristics of social learning models either (Supporting Figure S10). In stead, our simulation has suggested that self-organisation should play a key role to account for this emergent phenomenon. To seek a general mechanism underlying the observed collective behavioural rescue, we developed a reduced, approximated differential equation model that could give a time-depth qualitative insights into the collective decision-making dynamics.

2.3 The simplified collective dynamical model

In order to obtain qualitative understanding a synergy between positive and negative feedback that seems responsible for the pattern of adaptive behavioural shift observed in our individual-based simulation, we made a reduced model that approximates temporal changes of behaviour of an 'average' individual, where the computational details of reinforcement learning were purposely ignored. Specifically, we considered a differential equation model that focuses only on rises and falls of the expected number of individuals who chose the risky option and that of those who chose the safe option (Fig. 4a). Such dynamics modelling approach has been commonly used in population ecology and collective animal behaviour research, and has proven highly useful in disentangling factors underlying complex systems (e.g. Beckers et al., 1990; Goss et al., 1989; Seeley et al., 1991; Sumpter and Pratt, 2003; Camazine et al., 2001).

To confirm that this approximate model can successfully replicate the fundamental property of the hot stove effect, we first describe the asocial behavioural model without social influence (Materials and Methods). The baseline, asocial dynamical system has a locally stable non-trivial equilibrium that gives $N_S^\star \geq 0$ and $N_R^\star \geq 0$ where $N^\star$ means equilibrium density at which system stops changing ($dN_S^\star/dt = dN_R^\star/dt = 0$). At the equilibrium, a ratio between the number of individuals choosing the safe-option S and that choosing the risky-option R is given by $N_S^\star : N_R^\star = e(p_l/p_h) + (1-e)(p_h/p_l) : 1$, indicating that the risk-aversion (defined as a case where more people choose the safe option; $N_S^\star > N_R^\star$) emerges when the inequality $e < P_S^-/(P_S^- + P_R^-)$ holds.

Fig. 4b visually shows that the population is indeed attracted by the safe-option S, which results in $N_S^\star > N_R^\star$, in a wide range of the parameter region even when there is a positive 'risk premium' defined as $e > 1/2$. Although individuals choosing the risky-option R are more likely to get enchanted by the risky option (that is, to become R^+ state) than to get disappointed (that is, to become R^- state), the risk-seeking equilibrium, defined as $N_S^\star < N_R^\star$, could become harder to

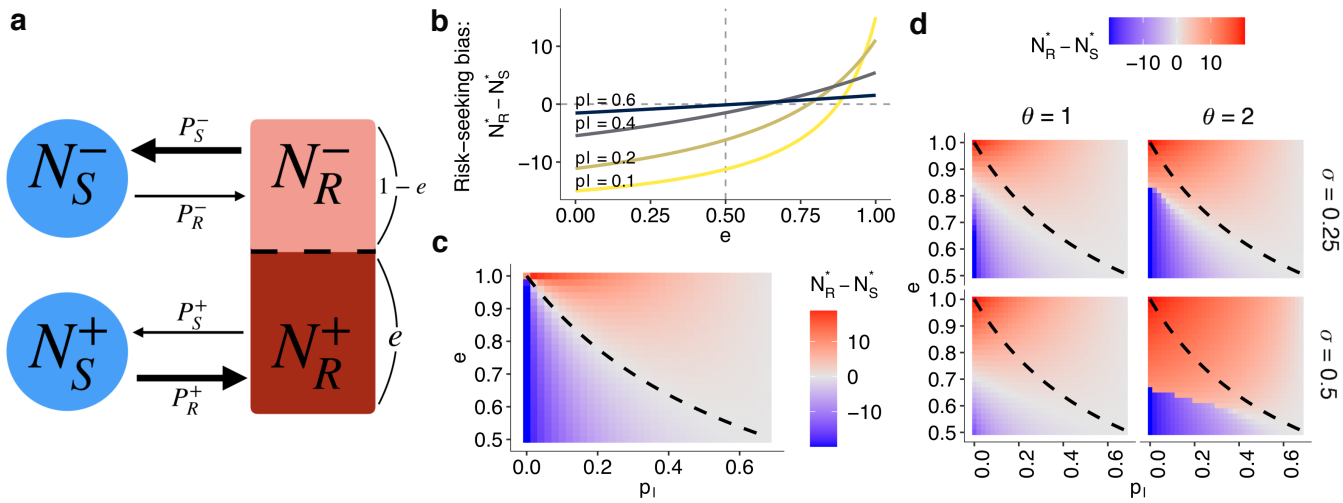

Figure 4: The collective behavioural dynamics model. (a) A schematic diagram of the dynamics. Solid arrows represent a change of population density between connected states at a time step. The thicker the arrow, the larger the per-capita rate of behavioural change. (b,c) The equilibrium bias toward risk-seeking (i.e. $N_r^\star - N_s^\star$) as a function of the degree of risk premium e as well as of the per-capita probability of moving to the less preferring behavioural option p_l. (b) The explicit form of the curve is given by $-n(p_h - p_l)\left\{(1-e)p_h - ep_l\right\}/(p_h + p_l)\left\{(1-e)p_h + ep_l\right\}$. (c,d) The equilibrium bias of risk-seeking ($N_R^* - N_S^* > 0$; red) or risk-aversion ($N_R^* - N_S^* < 0$; blue) depending on the combinations of p_l and e for (c) the asocial baseline dynamics and (d) the collective behavioural dynamics with social influences. The dashed curve is an analytical neutral equilibrium of the asocial system that results in $N_R^* = N_S^*$, given by $e = p_h/(p_h + p_l)$. The numerical results were obtained by $N_{S,t=0}^- = N_{S,t=0}^+ = 5$, $N_{R,t=0} = 10$, and $p_h = 0.7$

emerge with decreasing exploration rate p_l, consistent with the known property of asymmetric adaptive sampling that causes the hot stove effect (Denrell, 2007). The risk-seeking never emerges when $e \leq 1/2$, which is also consistent with the result of reinforcement learning.

This reduced dynamical model gives us an illustrative understanding how the asymmetry of adaptive sampling causes the hot stove effect. Consider a case of high inequality between exploitation (p_h) and exploration (p_l), namely $p_h \gg p_l$. Under such a condition, the behavioural state S^- becomes a 'dead-end' from which individuals could hardly escape once entered. However, if the inequality $p_h \geq p_l$ is not so large that the substantial fraction of population now come back to R^- from S^-, the number of people in R^+ could eventually overcome the amount of people 'spilling out' to S^-. Such an illustrative analysis depicts that the hot stove effect can be overcome if the number of people who get stacked in the dead-end S^- can be reduced. And this is possible if one can increase the 'come-back' to R^-. In other words, if any mechanisms can increase P_R^- in relation to P_S^-, the hot stove effect should be overcome.

Through numerical analyses of the dynamical model with social influence (see Materials and Methods), we confirmed that the social influence could indeed increase the flow-back rate P_R^-, which raised the possibility of risk-seeking equilibrium $N_R^\star > N_S^\star$ (Fig. 4d). For an approximation to the bifurcation analysis, we recorded an equilibrium density of the risky-state $N_R^\star$ starting from various initial distributions of both $N_{R,t=0}$ and $N_{S,t=0}$. Fig. 5 shows conditions under which the system ends up in the risk-seeking equilibrium. When the conformity exponent θ was not too large ($\theta < 10$), the risk-seeking can be a unique equilibrium irrespective of the initial distribution, attracting the population even from an extremely biased initial distribution such as $N_{R,t=0} = 0$ (Fig. 5).

Under the conformist-bias $\theta \geq 2$, a bifurcation with two locally-stable equilibria emerged. A strong positive feedback dominates the system when both σ and θ are large. Therefore, the system could end up in either of the equilibria depending solely on the initial density distribution. This is consistent with a well-known result of collective foraging by pheromone trail ants whose reaction to social information is a conformity-like manner (Beckers et al., 1990; Camazine et al., 2001).

Notably, however, even with a large conformity such as $\theta = 2$, there is a region of moderate value of σ where risk-seeking remains to be a unique equilibrium when the premium of risky-option e was high ($e \geq 0.7$). In this regime, the benefit of collective behavioural rescue could dominate without any cost of maladaptive herding.

It is worth noting that, in the case of $\theta = 0$ where individuals made merely a random choice at a rate σ, the risk aversion was also relaxed (Fig. 5), and the adaptive risky-shift even emerges around $0.25 < \sigma < 1$. However, this ostensible behavioural rescue was solely due to the pure effect of additional random exploration that reduces $P_S^-/(P_S^- + P_R^-)$, mitigating stickiness to the dead-end. When $\sigma \approx 1$, therefore, the risky-shift disappeared because population choose between S and R almost randomly.

However, the collective risky-shift observed in the conditions of $\theta > 0$ cannot be explained solely by the mere addition of exploration. A weak conformist-bias (i.e. a linear response to the social frequency; $\theta = 1$) monotonically increases the equilibrium density $N_R^\star$ with increasing social influence σ, which goes beyond the level of risky-shift observed by the additional random choice (Fig. 5). Therefore, although the collective rescue might indeed owe its part to the effect of exploration that reduces the ratio $P_S^-/(P_S^- + P_R^-)$, it does not fully explain the further enhancement of risk-seeking.

The key is an interaction between negative and positive feedback. As we discussed above, the risk-aversion is reduced if the ratio $P_S^-/(P_S^- + P_R^-)$ decreases either by increasing P_R^- or reducing P_S^-. The per-capita probability choosing the safe option with belief $-$, that is $P_S^- = (1-\sigma)p_h + \sigma N_S^\theta/(N_R^\theta + N_S^\theta)$, becomes smaller than the baseline exploitation

probability p_h when $N_S^\theta/(N_R^\theta + N_S^\theta) < p_h$. Even though the majority of the population might still choose the safe-alternative and hence $N_S > N_R$ holds, the inequality $N_S^\theta/(N_R^\theta + N_S^\theta) < p_h$ can nevertheless hold with a sufficiently small value of θ. Crucially, the reduction of P_S^- leads further reduction of P_S^- itself through decreasing N_S^-, and thereby further decreasing the social influence supporting the safe option $N_S^\theta/(N_R^\theta + N_S^\theta)$. Such a process creates a negative feedback that weakens the concomitant risk-aversion. Naturally, this negative feedback could be maximised with $\theta = 0$.

Once the negative feedback has weakened the underlying risk-aversion, the majority of the population may eventually choose the risky option, an effect evident in the case $\theta = 0$ (Fig. 5). What uniquely operates in cases of $\theta > 0$ is that, because N_S is minority by now, here starts a positive feedback. Thanks to the conformity bias, the inequality $N_R > N_S$ is further *amplified*. In this phase, the larger θ the stronger the concomitant relationship $N_S^\theta/(N_R^\theta + N_S^\theta) \ll p_h$. Such a positive feedback could never operate with $\theta = 0$.

In conclusion, it is a synergy between negative and positive feedback that explains the full range of adaptive risky-shift. Neither positive nor negative feedback alone could account for both accuracy and flexibility emerging through collective learning and decision-making. The results are qualitatively unchanged across a range of different combinations of e, p_l, and p_h (Supporting Figure 13 and 14). It is worth noting that when $e < 0.5$, the system goes to the risk-aversion equilibrium (Supporting Figure 14), consistent with the result of agent-based simulation for the case where the mean payoff of the risky option was smaller than that of the safe option (Supporting Figure 11). Therefore, the system does not mindlessly prefers risk-seeking, but it prefers risk-seeking only when to do so is optimal.

2.4 An experimental demonstration

A series of online behavioural experiments confirmed that collective learning could indeed mitigate the suboptimal risk-aversion. One hundred sixty adult human subjects performed the individual task without social interactions, while 307 subjects performed the task collectively with group size ranging from 3 to 8 (Supporting Fig. 16). We used three different settings for the multi-armed bandit task (see Materials and Methods).

Fig. 6 shows that participants who had a larger estimated value of $\alpha_i(\beta_i + 1)$ were less likely to choose the optimal risky alternative, whereas those who had a smaller value of $\alpha_i(\beta_i + 1)$ had a higher chance to escape from suboptimal risk-aversion in both two-armed and four-armed tasks, consistent with the theory of hot stove effect (Fig. 2). This suggests that the parameter estimation successfully captured participants' behavioural pattern.

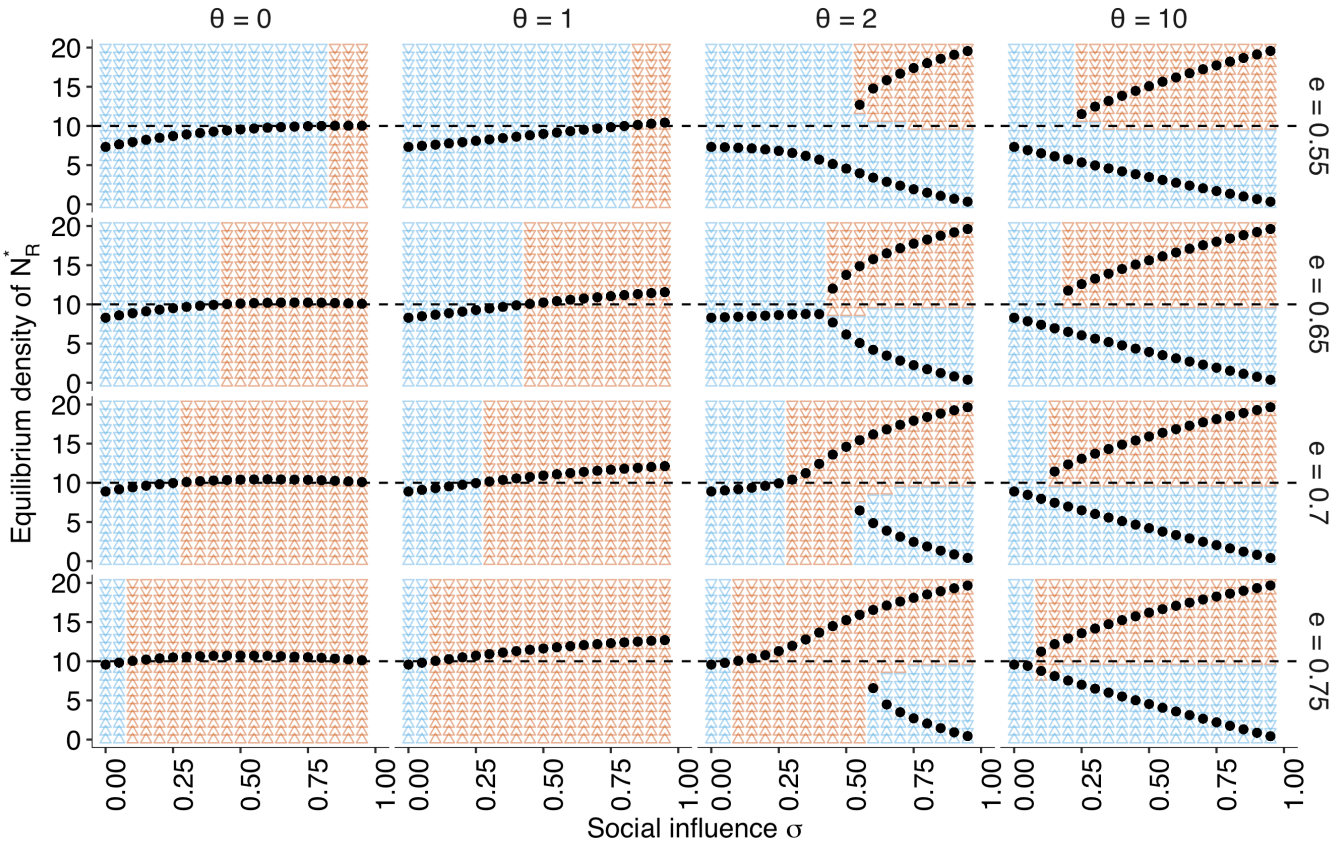

Figure 5: The approximate bifurcation analysis. The relationships between the social influence weight σ and the equilibrium number of the individuals in the risky behavioural state $N_R^\star$ across different conformity exponent $\theta \in \{0, 1, 2, 10\}$ and different degrees of risk premium $e \in \{0.55, 0.65, 0.7, 0.75\}$, are shown in the black dots. The colour in background indicates regions where the system approaches either to risk-aversion ($N_R^\star < N_S^\star$; blue) or to risk-seeking ($N_R^\star > N_S^\star$; red). The horizontal dashed line is $N_R = N_S = 10$. Two locally-stable equilibria emerges when $\theta = 2$ and $\theta = 10$, which suggests that the system has a bifurcation when σ is sufficiently large. The other parameters are set to $p_h = 0.7$, $p_l = 0.2$, and $N = 20$.

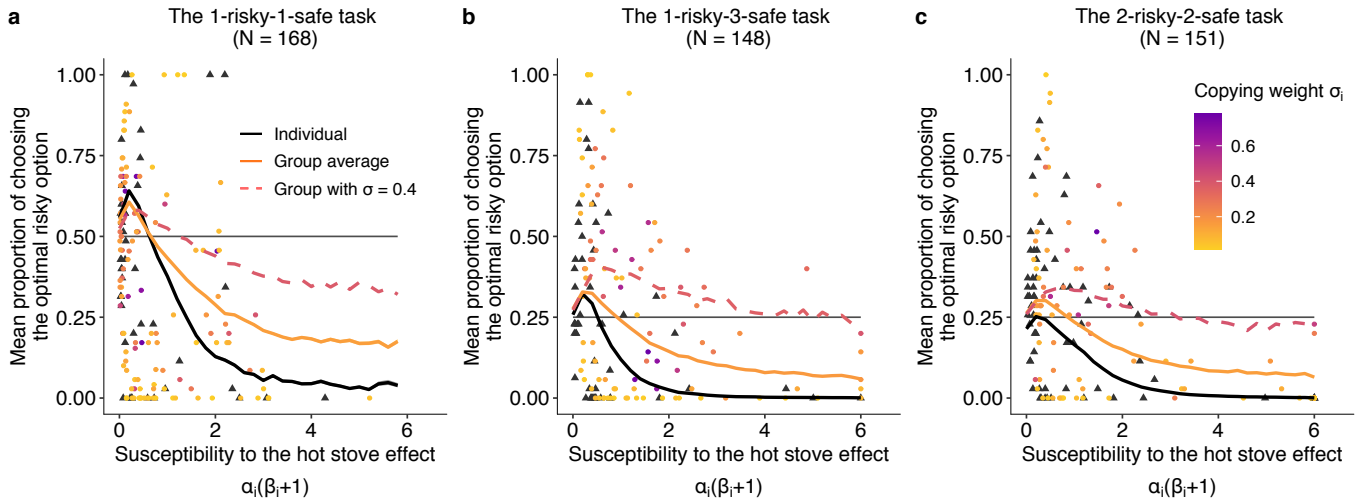

Figure 6: The effect of social learning on the decision performance in the online behavioural experiments. (a) a two-armed bandit task ($N = 168$), (b) an 1-risky-3-safe (4-armed) bandit task ($N = 148$), and (c) a 2-risky-2-safe (4-armed) bandit task ($N = 151$). The x-axis is $\alpha_i(\beta_i + 1)$, namely, the susceptibility to the hot stove effect. The y-axis is the mean proportion of choosing the optimal risky alternative averaged over the second half of the trials. The black triangular-dots represent subjects in the individual learning condition, whereas the coloured dots are those in the group condition with group size ranging between 4 to 8. For subjects in the group condition, the fit value of copying weight σ_i was shown in colour gradient. The solid lines are predictions from the model with the calibrated global parameters for the individual condition (black) and for the group condition (orange). The dashed violet curves are hypothetical prediction of the calibrated model supplied with a larger copying weight than real human subjects, that is $\overline{\sigma}_i = 0.4$, showing a larger rescue effect potentially emerging with a heavier social influence.

The behaviour of the group condition supported our theoretical predictions. The proportion of choosing the optimal risky-option increased by social influence (σ_i) particularly for individuals who had a large susceptibility to the hot stove effect. On the other hand, social influence had little aid for those who had a small susceptibility to the hot stove effect (e.g. $\alpha_i(\beta_i + 1) \leq 0.5$), which also supported our theory.

However, a complete alternation of the majority's behaviour from the suboptimal options to the optimal risky option (that is, $P_r > 0.5$ for the 2-armed task and $P_r > 0.25$ for the 4-armed task) was not widely observed (Supporting Fig. 16). This was due to the low reliance on social learning (that is, the mean $\sigma_i = 0.19$ and the median $= 0.12$), coupled with the small susceptibility $\alpha_i(\beta_i + 1)$ in individual learners (that is, the mean $\alpha_i(\beta_i + 1) = 0.8$ and median $= 0.3$ for the individual condition, whereas the mean $= 1.1$ and median $= 0.5$ for the group condition). Although the mitigation of risk-aversion still emerged for those with higher $\alpha_i(\beta_i + 1)$ (Fig. 6), the weak reliance on social learning hindered the strong collective rescue effect because strong positive feedback was not robustly formed.

The estimated global mean of the conformity exponent μ_θ was larger than 1 for all three tasks (Table 1), suggesting that human groups might still have possessed a potential to set out positive-feedback loops required for the strong rescue effect. Consistently, if we extrapolated the larger value of copying weight (that is, $\bar{\sigma}_i = 0.4$) into the best-fit social learning model with calibrated other parameters, the strong collective rescue became prominent (Fig. 6, the violet dashed lines).

Table 1: The mean and the 95% Bayesian credible intervals of the global parameters. All parameters satisfied the Gelman–Rubin criterion $\hat{R} < 1.01$. All estimates are based on over 500 effective samples from the posterior.

Tasks:	1-risky-1-safe	1-risky-3-safe	2-risky-2-safe
Group condition	n = 123	n = 97	n = 87
$\mu_{\text{logit}\alpha}$	-2.2 [-2.8, -1.5]	-1.8 [-2.3, -1.4]	-1.7 [-2.1, -1.3]
(Mean α)	0.10 [0.06, 0.18]	0.14 [0.09, 0.20]	0.15 [0.11, 0.21]
$\mu_{\log\beta}$	1.4 [1.1, 1.6]	1.5 [1.3, 1.8]	1.3 [1.0, 1.5]
(Mean β)	4.1 [3.0, 5.0]	4.5 [3.7, 6.0]	3.7 [2.7, 4.5]
$\mu_{\text{logit}\sigma}$	-2.4 [-3.1, -1.8]	-2.1 [-2.6, -1.6]	-2.1 [-2.5, -1.7]
(Mean σ)	0.08 [0.04, 0.14]	0.11 [0.07, 0.17]	0.11 [0.08, 0.15]
μ_θ	1.4 [0.58, 2.3]	1.6 [0.9, 2.4]	1.8 [1.0, 2.9]
Individual condition	n = 45	n = 51	n = 64
$\mu_{\text{logit}\alpha}$	-2.1 [-3.1, -0.87]	-2.1 [-2.6, -1.6]	-1.3 [-2.1, -0.50]
(Mean α)	0.11 [0.04, 0.30]	0.11 [0.07, 0.17]	0.21 [0.11, 0.38]
$\mu_{\log\beta}$	0.42 [-0.43, 1.1]	0.91 [0.63, 1.2]	0.76 [0.42, 1.1]
(Mean β)	1.5 [0.65, 3.0]	2.5 [1.9, 3.3]	2.1 [1.5, 3.0]

3 Discussion

We have demonstrated that frequency-based copying, one of the commonest form of social learning strategies, can rescue decision-makers from committing to suboptimal risk-aversion in a risky trial-and-error learning task, even though a majority of individuals is potentially biased toward suboptimal risk-aversion. Although an extremely strong reliance on the conformist-copying can raise a possibility of getting stuck to the suboptimal option, the mitigation of risk-aversion, and the concomitant collective behavioural rescue, could emerge in a wide range of situations under the modest use of conformist-based social learning.

Our simplified differential equation model has identified a key mechanism of the collective behavioural rescue: a synergy between positive and negative feedback. Despite conformity the probability of choosing the suboptimal option can decrease from what expected by individual learning alone. Indeed, an inherent individual preference to the safe alternative (that is, the softmax $e^{\beta Q_s}/(e^{\beta Q_s} + e^{\beta Q_r})$) is always mitigated by the conformist-influence $N_s^\theta/(N_s^\theta + N_r^\theta)$ if the former is larger than the latter. Once the individual risk-aversion is reduced, there will exist fewer individuals choosing the suboptimal safe option, which further weakens the majority support toward the safe option. This negative feedback facilitates individuals to revisit the risky alternative. Such an attraction toward the risky option allows more individuals, including those who are currently sceptical about the value of the risky option, to experience a large bonanza from the risky option, which results in 'gluing' them to the risky alternative for a while. Once a majority of individuals get glued in the risky alternative, a positive feedback by conformity kicks off, and the optimal risk-seeking is further strengthened.

The temporally flexible collective behaviour emerging under the positive frequency-biased copying has repeatedly been suggested by previous empirical studies (Toyokawa et al., 2019; Aplin et al., 2017; Beckers et al., 1990; Seeley et al., 1991; Camazine et al., 2001; Kandler and Laland, 2013). Our model can give a general explanation to these results too. For example, Aplin et al. (2017) demonstrated that populations of the great tits (*Parus major*) could switch their behavioural tradition after an environmental change even though individual birds were likely to have a strong conformist tendency. A similar phenomenon was also reported in humans (Toyokawa et al., 2019). Although these studies did not focus on a risky decision-making, and hence individuals were not inherently biased, an experimentally-induced environmental change could create such a situation where majority individuals exhibited a suboptimal behaviour. Many individuals in a population were choosing the out-dated suboptimal behavioural option soon after the environmental change. However, as we have shown, the collective learning system can recover their performance flexibly even though the individual distribution was strongly biased toward the suboptimal direction at the outset. The great tit and human groups were able to switch their tradition because of, rather than despite of, the conformist-biased social influence, thanks to the synergy of both negative and positive feedback processes.

Through the online behavioural experiments using a risky multi-armed bandit task, we have confirmed our theoretical prediction that the simple frequency-based copying could reduce suboptimal risk-aversion that many individual learners, especially those who had higher learning rates and/or lower exploration rates, would have exhibited due to the hot stove effect. Yet, there still was a majority, albeit a smaller one, who showed risk aversion. The weak reliance on social learning which affected only about 15% of decision-making, was unable to facilitate a strong positive feedback.

The little use of social information might be due to the stationarity of the task. In a stable environment, learners could eventually gather enough information as trials proceeded, which might make learners less curious on information gathering including social learning (Rendell et al., 2010). Future research will explore for conditions under which humans elicit heavier reliance on social learning, such as volatility of environment, time-pressure, or increasing number of behavioural options (Muthukrishna et al., 2015), coupled with larger group sizes (Toyokawa et al., 2019).

The low learning rate α, which was at most 0.2 for many individuals, should also have hindered potential benefits of collective rescue in our current experiment, because benefit of mitigating the hot stove effect would be minimum or hardly realise under such a small susceptibility to the hot stove effect. Although we believe that the simplest stationary environment in our current analyses and experiments were a necessary first step in building our understanding of the collective behavioural rescue effect, we would suggest for the future study to use a temporally unstable ('restless') bandit task to elicit both a higher learning rate as well as a heavier reliance on social learning, so as to investigate the possibilities of stronger effect . Indeed, previous studies with changing environments have reported a learning rate as high as $\alpha > 0.5$ (Toyokawa et al., 2017, 2019; Deffner et al., 2020), under which individual learners should have suffered the hot stove trap.

Information about others' payoff might also be available in addition to inadvertent social frequency cues in some cooperative contexts, especially with an aid of benevolent pedagogical instructions by others. Although communicative acts may transfers information about unexplored options and forgone payoffs, which potentially mitigates the hot stove effect (Yechiam and Busemeyer, 2006), it may further amplify the suboptimal decision bias if instructors, albeit their cooperative motivations, selectively filters out some pieces of information that they thought were redundant (Moussaïd et al., 2015). To the contrary, previous studies suggest that competitions or conflict of interests among individuals can lead better collective intelligence than fully cooperative situations (Conradt et al., 2013), and can promote adaptive risk-taking (Arbilly et al., 2011). Further research will identify conditions under which cooperative communication can improve decision-making, and drive adaptive cumulative cultural evolution (Csibra and Gergely, 2011; Morgan et al., 2015), when suboptimal individual behavioural biases prevail.

Not only human groups, the generality of our dynamical model should apply to various collective decision-making systems. Because it is a fundamental property of adaptive reinforcement learning, risk-aversion due to the hot stove effect should be widespread in animals (Real, 1981; Weber et al., 2004; Hertwig and Erev, 2009). Therefore, its solution, the collective behavioural rescue, should also operate broadly in collective animal decision-making because the frequency-based copying is one of the common social learning strategies (Hoppitt and Laland, 2013; Grüter and

Leadbeater, 2014). Future research should determine to what extent the collective behavioural rescue actually impacts on animal decision-making in wider contexts, and whether it influences on the evolution of social learning, information sharing, and the formation of group-living.

We have identified a previously overlooked mechanism underlying adaptive advantages of the frequency-based social learning. Our results suggest that an informational benefit of group-living could exist well beyond a simple informational pooling where individuals can enjoy the wisdom of crowds effect (Ward and Zahavi, 1973). Furthermore, the flexibility emerging through the interaction of negative and positive feedback suggests that conformity could evolve in a wider range of environments than previously assumed (Aoki and Feldman, 2014; Nakahashi et al., 2012), including temporally variable environments (Aplin et al., 2017). Social learning can drive self-organisation regulating the mitigation and amplification of behavioural biases, canalising the course of repeated decision-making under risk and uncertainty.

4 Methods

4.1 The baseline learning model and hot stove effect

We considered a decision-maker facing a repeated choice between a safe (i.e. certain) alternative s that generated a constant payoff π_s, and a risky (i.e. uncertain) alternative r that generated a random payoff π_r following a Gaussian distribution $\mathcal{N}(\mu, \text{s.d.})$ (i.e. a two-armed risky bandit task; Fig. 1a).

We assumed that a decision-maker updates their value of choosing the alternative i ($\in \{s, r\}$) at time t, following the Rescorla-Wagner learning rule: $Q_{i,t+1} \leftarrow (1-\alpha)Q_{i,t} + \alpha\pi_{i,t}$ where α ($0 \leq \alpha \leq 1$) is a *learning rate*, manipulating a step size of the belief updating, and $\pi_{i,t}$ is a realised payoff from the chosen alternative i at time t (Sutton and Barto, 2018). The larger the α, the more weight is given to recent experiences, making reinforcement learning more myopic. The Q value for the unchosen alternative is unchanged. Before the first choice, individuals had no previous preference for either option (that is, $Q_{r,1} = Q_{s,1} = 0$). Then Q values are translated into choice probabilities through a softmax (or multinomial-logistic) function such that $P_{i,t} = \exp(\beta Q_{i,t})/(\exp(\beta Q_{s,t}) + \exp(\beta Q_{r,t}))$ where β, an *inverse temperature*, is a parameter regulating how sensitive the choice probability is to the value of the estimate Q (i.e. controlling the proneness to explore). As $\beta \to 0$, the softmax probability approximates to a random choice (i.e. highly explorative). Conversely, if $\beta \to +\infty$, it assymptotes to a deterministic choice in favour of the option with the highest Q value (that is, highly exploitative).

In this simplest setup, Denrell (2007) formally derived a condition under which a decision-maker exhibits risk-aversion. When the number of decision trials is infinite ($t \to \infty$), risk-aversion emerges under $\beta > (2-\alpha)/\alpha$, which corresponds to the area above the dashed curve in Figure 1. See Supplementary Methods for the details of the analytical results.

4.2 Collective learning and social influences

Based on the simplest setup of the two-armed bandit (March, 1996; Denrell, 2007), we considered a collective learning situation in which a group of individuals play the task simultaneously where individuals can obtain social information. For social information, we assumed a simple frequency-based social cue specifying distributions of individual choices (McElreath et al., 2005, 2008; Toyokawa et al., 2017, 2019; Deffner et al., 2020). We assumed that individuals cannot observe others' earnings, ensuring that individuals cannot learn from approximate forgone payoffs that would have easily rescued people from the hot stove effect (Denrell, 2007; Yechiam and Busemeyer, 2006).

A payoff realised was independent of others' decisions and it was drawn solely from the payoff probability distribution specific to each alternative, meaning that there is no direct social competition over the monetary reward. We assumed that individuals knew this was the case, and hence we ignored possibilities of the competitive motivations in decision-making.

Following the previous modelling of social learning in such multi-agent multi-armed bandit situations (e.g. Aplin et al., 2017; Barrett et al., 2017; McElreath et al., 2005, 2008; Toyokawa et al., 2017, 2019; Deffner et al., 2020), we assumed that social influences on reinforcement learning was expressed as a weighted average between the softmax probability based on the Q values and the conformist social influence as follows:

$$P_{i,t} = (1-\sigma)\frac{\exp(\beta Q_{i,t})}{\exp(\beta Q_{r,t}) + \exp(\beta Q_{s,t})} + \sigma\frac{(N_{i,t-1}+0.1)^{\theta}}{(N_{s,t-1}+0.1)^{\theta} + (N_{r,t-1}+0.1)^{\theta}} \tag{1}$$

where σ was a weight given to the social influence (*copying weight*) and θ was the strength of conformist influence (*conformity exponent*) which determines the influence of social frequency of choosing the alternative i at time $t-1$, that is $N_{i,t-1}$. The larger the conformity exponent θ, the higher influence was given to an alternative that was chosen by more individuals, with non-linear conformist-biased social influence arising when $\theta > 1$. Note that there was no actual social influence when $\theta = 0$ because in this case the 'social influence' favours a uniformly random choice, irrespective of the social frequency distribution N_s and N_r. We added a small number 0.1 to $N_{i,t-1}$ so that an option chosen by no one (that is, $N_{i,t-1} = 0$) could provide the highest social influence when $\theta < 0$ (negative frequency bias). Although this additional 0.1 slightly reduces the conformity influence when $\theta > 0$, we confirmed that the results were qualitatively

Table 2: Summary of the learning model parameters

Symbol	Meaning	Range of the value
α	Learning rate	$[0, 1]$
β	Inverse temperature	$[0, +\infty]$
$\alpha(1+\beta)$	Susceptibility to the hot stove effect	
σ	Copying weight	$[0, 1]$
θ	Conformity exponent	$[-\infty, +\infty]$

unchanged. Note also that, in the first trial $t = 1$, we assumed that the choice was determined solely by the asocial softmax function because there is no social information available yet. The descriptions of the parameters are shown in Table 2.

5 The approximated dynamical model of collective behaviour

We assume a group of N individuals consisting of two different behavioural states: one choosing a safe alternative S that contains N_S individuals; and the other choosing a risky alternative R that contains N_R individuals ($N = N_S + N_R$). We also assume that there are two different 'inner belief' states, labelled '$-$' and '$+$'. Individuals who possess the negative belief, labelled $-$, prefer the safe alternative S to R, while those who possess the positive belief $+$ prefer R to S. A per-capita probability of choice shift from one behavioural alternative to the other is denoted by P. For example, P_S^- means an indivdual probability of changing the choice to the safe alternative from the risky alternative under the negative belief. Because there exist N_S^- individuals who chose S with belief $-$, a total number of individuals who 'move out' to S from R at one time step is denoted by $P_S^- N_S^-$. We assume that the probability of shifting to the more preferable option is larger than that of shifting to less preferable option, that is $P_S^- > P_R^-$ and $P_R^+ > P_S^+$ (Figure 4a).

We assume that the belief state can change by choosing the risky alternative. We define that the per-capita probability of becoming $+$ state, that is having a higher preference to the risky alternative, is e ($0 \leq e \leq 1$), and hence $N_R^+ = eN_R$. The rest of individuals who choose the risky alternative become a belief state $-$, that is, $N_R^- = (1-e)N_R$.

We define e so that it can be seen as a risk premium of gambles. For example, imagine a 2-armed bandit task equipped with one risky arm with Gaussian noises and the other sure arm. The larger the mean expected reward of the risky option (i.e. higher risk premium), the more people who choose the risky arm are expected to obtain a larger reward

than what the safe alternative would provide. By assuming $e > 1/2$, therefore, it approximates a situation where risk-seeking is optimal.

Here we focus only on the population dynamics: If more people choose S, N_S increases. On the other hand, if more people choose R, N_R increases. As a consequence, the system may eventually reach an equilibrium state where both N_S and N_R do not change any more. We analyse such a case, and if the equilibrium state satisfies $N_R^\star > N_S^\star$, we define that the population exhibit risk-seeking, escaping from the hot stove effect.

In sum, the dynamics are written as the following differential equations:

$$\begin{cases} \frac{dN_R}{dt} = P_R^- N_S^- - P_S^- N_R^- + P_R^+ N_S^+ - P_S^+ N_R^+ \\ \frac{dN_S^-}{dt} = -P_R^- N_S^- + P_S^- N_R^-, \\ \frac{dN_S^+}{dt} = -P_R^+ N_S^+ + P_S^+ N_R^+ \end{cases} \tag{2}$$

where $N_R^+ = eN_R$ and $N_R^- = (1-e)N_R$. For the sake of simplicity, we assume $p_l = P_R^- = P_S^+$ and $p_h = P_R^+ = P_S^-$ where $0 \leq p_l \leq p_h \leq 1$. The equilibrium analysis was conducted in Mathematica (code is available online).

Overall, our model crystallises the asymmetry emerging from adaptive sampling which is considered as a fundamental mechanism of the hot stove effect (Denrell, 2007; March, 1996): Once decision-makers underestimate the expected value of the risky alternative, they start avoiding it and cannot have another chance to correct the error. Although there would potentially be more individuals who possess a positive preference to the risky alternative (i.e. $+$ belief), this asymmetry may ratchet-up the number of people who choose the safe alternative with underestimation of the value of the risky alternative (i.e. $-$ belief). If our model is able to capture this asymmetrical dynamics properly, the relationship between e (i.e. the potential goodness of the risky option) versus p_l and p_h (i.e. the exploration-exploitation) should account for the hot stove effect as suggested by previous learning model analyses (Denrell, 2007).

5.1 Collective dynamics with social influences

For approximation of the social influences, we assumed that the behavioural transition rates, S and R, would depend on the number of individuals N_S and N_R as follows:

$$\begin{cases} P_S^- = (1-\sigma)p_h + \sigma\frac{N_S^\theta}{N_R^\theta+N_S^\theta}, \\ P_R^- = (1-\sigma)p_l + \sigma\frac{N_R^\theta}{N_R^\theta+N_S^\theta}, \\ P_S^+ = (1-\sigma)p_l + \sigma\frac{N_S^\theta}{N_R^\theta+N_S^\theta}, \\ P_R^+ = (1-\sigma)p_h + \sigma\frac{N_R^\theta}{N_R^\theta+N_S^\theta}, \end{cases} \tag{3}$$

where σ is a weight of social influence and θ is a strength of conformist bias, corresponding the agent-based learning model. Other assumptions were the same as in the baseline dynamics model. The baseline dynamical model was a special case of the social influence model where $\sigma = 0$. Because the system was not analytically tractable, we obtained the numeric solution across different initial distribution of N_S and N_R for various combinations of the parameters.

5.2 The online experiments

The experimental procedure was approved by the Ethics Committee at the University of Konstanz ("Collective learning and decision-making study"). English-speaking 492 subjects (234 self-identified as women, 211 as men, and 1 as others; mean (minimum, maximum) age = 34.8 (18, 74) years) participated in the task through an online experimental recruiting platform Prolific Academic. We excluded subjects who disconnected from the online task before completing at least the first 35 rounds from our computational model-fitting analysis, resulting in 467 subjects (the detailed distribution of subjects for each condition is shown in Table 1). A parameter recovery test had suggested that the sample size was sufficient to reliably estimate individual parameters using a hierarchical Bayesian fitting method (see below; Supplementary Fig. S15).

5.2.1 Design of the experimental manipulations

The group size was manipulated by randomly assigning different capacities of the 'waiting lobby' where participants had to wait other participants arrival. When the lobby capacity was 1, which happened at probability 0.1, the individual condition started upon the first subject's arrival. Otherwise, the group condition started when there were more than 3 people at 3 minutes being past since the lobby opened (see Online Supplementary method). If there were only less than or equal to 3 people in a lobby at this stage, the participants each were assigned to the individual condition (i.e. all groups were larger than 3). The distribution of group size is shown in Supporting Fig. 16.

We used 3 different tasks which were 1-risky-1-safe task, 1-risky-3-safe task, and 2-risky-2-safe task. All tasks had 70 decision-making trials. The task proceeded at a trial-basis, that is, trials of all individuals in a group were synchronised.

Participants in the group condition could see the social frequency information, namely, how many people chose each slot in the preceding trial. No social information was available in the first trial. Each task setting was randomly assigned for each different session, and participants were allowed to participate in one session only.

We employed a skewed payoff probability distribution rather than a normal distribution for the risky alternative, as well as conducting not only a 2-armed task but also 4-armed bandit tasks, because our pilot study had suggested that participants tended to have a small susceptibility to the effect ($\alpha_i(\beta_i + 1) \ll 2$), hence we needed more difficult settings than the conventional Gaussian noise binary-choice task to elicit risk-aversion from individual decision-makers.

The details of the 3 task setups were as follows:

The 1-risky-1-safe task. The optimal risky option produced either 50 or 550 points at probability 0.7 and 0.3, respectively (the expected payoff was 200). The safe option produced 150 points (with a small Gaussian noise with s.d. = 5).

The 1-risky-3-safe task. The optimal risky option produced either 50 or 425 points at probability 0.6 and 0.4, respectively (the expected payoff was 200). The 3 safe options each produced 150, 125, and 100 points respectively, with a small Gaussian noise with s.d. = 5.

The 2-risky-2-safe task. The optimal risky option produced either 50 or 425 points at probability 0.6 and 0.4, respectively (the expected payoff = 200). The two safe options each produced 150 and 125 points, respectively, with a small Gaussian noise with s.d. = 5. The suboptimal risky option whose expected value was 125, produced either 50 or 238 points at probability 0.6 and 0.4, respectively.

We have confirmed that the collective behavioural rescue could emerge in a task equipped with a Binary noise of the risky option (Supplementary Fig. S12). The use of the 4-armed tasks were aimed to demonstrate that the rescue effect is not limited to the binary-choice situations. Other procedures of collective learning task were the same as the task used in our agent-based simulation. The experimental materials including illustrated instructions can be found in Supplementary Video SV1 (individual condition) and SV2 (group condition).

5.2.2 The hierarchical Bayesian model fitting

To fit the learning model, we used a hierarchical Bayesian method, estimating the global means (μ_α, μ_β, μ_σ, and μ_θ) and global variances (v_α, v_β, v_σ, and v_θ) for each of the three experimental conditions and for the individual and group conditions separately. For the individual condition, we assume $\sigma = 0$ for all subjects and hence no social

learning parameters were estimated. Full details of the model-fitting procedure and prior assumptions are shown in the Supplementary method. The R and Stan code used in the model fitting are available from an online repository.

Parameter recovery and post-hoc simulation To assess the adequacy of the hierarchical Bayesian model-fitting method, we tested how well the HBM could recover 'true' parameter values that were used to simulate synthetic data. We simulated artificial agents' behaviour assuming that they behave according to the social learning model with each parameter setting. We generated 'true' parameter values for each simulated agent based both on the experimentally fit global parameters (Table 1 in the main text; parameter recovery test 1). In addition, we ran another recovery test using arbitrary global parameters that deviated from the experimental fit values (parameter recovery test 2), so as to confirm that our fitting procedure was not just 'attracted' toward the fit value. We then simulated synthetic behavioural data and recovered their parameter values using the HBM described above. The parameter recovery test showed that all the recovered individual parameters were positively correlated with the true values, whose correlation coefficients were all larger than 0.5. We also confirmed that 30 out of 32 global parameters in total were recovered within the 95% Bayesian credible intervals, and that even those two non-recovered parameters (μ_β for the 2-risky-2-safe task in the recovery test 1 and for μ_α for the 1-risky-3-safe task in the recovery test 2) were not so deviated from the true value (Supplementary Fig. 15).

We also ran a series of individual-based model simulations using the calibrated global parameter values for each condition. First, we randomly sampled a set of agents whose individual parameter values were drawn from the fit global parameters. Second, we let this synthetic group of agents perform the task for 70 rounds. We repeated these steps 100,000 times for each task setting and for each individual and group condition.

6 Data availability

Both experimental and simulation data are available in an online repository (ttps://github.com/WataruToyokawa/ToyokawaGaissmaier2021)

References

S. Alem, C. J. Perry, X. Zhu, O. J. Loukola, T. Ingraham, E. Søvik, and L. Chittka. Associative Mechanisms Allow for Social Learning and Cultural Transmission of String Pulling in an Insect. *PLoS Biology*, 14(10):1–28, 2016. ISSN 15457885. doi: 10.1371/journal.pbio.1002564.

K. Aoki and M. W. Feldman. Evolution of learning strategies in temporally and spatially variable environments: a review of theory. *Theoretical population biology*, 91:3–19, feb 2014. ISSN 1096-0325. doi: 10.1016/j.tpb.2013.10.004. URL `http://www.ncbi.nlm.nih.gov/pubmed/24211681`.

L. M. Aplin, B. C. Sheldon, and R. McElreath. Conformity does not perpetuate suboptimal traditions in a wild population of songbirds. *Proceedings of the National Academy of Sciences*, page 201621067, 2017. ISSN 0027-8424. doi: 10.1073/PNAS.1621067114. URL `http://www.pnas.org/content/early/2017/07/18/1621067114.full`.

M. Arbilly, U. Motro, M. W. Feldman, and A. Lotem. Evolution of social learning when high expected payoffs are associated with high risk of failure. *Journal of the Royal Society Interface*, 8(64):1604–1615, 2011. ISSN 17425662. doi: 10.1098/rsif.2011.0138.

B. J. Barrett, R. L. Mcelreath, S. E. Perry, and B. J. Barrett. Pay-off-biased social learning underlies the diffusion of novel extractive foraging traditions in a wild primate. *Proceedings of Royal Society B*, 284, 2017.

R. Beckers, J. L. D. Deneubourg, S. Goss, and J. M. Pasteels. Collective decision making through food recruitment. *Insectes Sociaux*, 37:258–267, 1990.

G. Biele, J. Rieskamp, L. K. Krugel, and H. R. Heekeren. The Neural basis of following advice. *PLoS Biology*, 9(6), 2011. ISSN 15449173. doi: 10.1371/journal.pbio.1001089.

R. Boyd and P. J. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, Chicago, IL, 1985.

S. Camazine, J.-L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau. *Self-Organization in Biological Systems*. Princeton University Press, Princeton, NJ, 2001. ISBN 9780691116242.

D. Chung, G. I. Christopoulos, B. King-Casas, S. B. Ball, and P. H. Chiu. Social signals of safety and risk confer utility and have asymmetric effects on observers' choices. *Nature Neuroscience*, 18(6):3–6, 2015. ISSN 1097-6256. doi: 10.1038/nn.4022. URL `http://www.nature.com/doifinder/10.1038/nn.4022`.

L. Conradt, C. List, and T. J. Roper. Swarm Intelligence: When Uncertainty Meets Conflict. *The American Naturalist*, pages 000–000, aug 2013. ISSN 00030147. doi: 10.1086/673253. URL `http://www.jstor.org/stable/info/10.1086/673253`.

G. Csibra and G. Gergely. Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567):1149–1157, 2011. ISSN 14712970. doi: 10.1098/rstb.2010.0319.

D. Deffner, V. Kleinow, and R. McElreath. Dynamic Social Learning in Temporally and Spatially Variable Environments. *Royal Society Open Science*, 7:200734, 2020. doi: 10.1098/rsos.200734.

J. Denrell. Adaptive learning and risk taking. *Psychological Review*, 114(1):177–187, 2007. ISSN 0033295X. doi: 10.1037/0033-295X.114.1.177.

J. Denrell and J. G. March. Adaptation as Information Restriction: The Hot Stove Effect. *Organization Science*, 12(5): 523–538, 2001. ISSN 10477039. doi: 10.1287/orsc.12.5.523.10092.

T. Drezner-Levy and S. Shafir. Parameters of variable reward distributions that affect risk sensitivity of honey bees. *Journal of Experimental Biology*, 210(2):269–277, 2007. ISSN 00220949. doi: 10.1242/jeb.02656.

A. Dussutour, J. L. Deneubourg, and V. Fourcassié. Amplification of individual preferences in a social context: The case of wall-following in ants. *Proceedings of the Royal Society B: Biological Sciences*, 272(1564):705–714, 2005. ISSN 14712970. doi: 10.1098/rspb.2004.2990.

R. Frey, A. Pedroni, R. Mata, J. Rieskamp, and R. Hertwig. Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, 3(10):1–14, 2017. ISSN 23752548. doi: 10.1126/sciadv.1701381.

L.-A. Giraldeau, T. J. Valone, and J. J. Templeton. Potential disadvantages of using socially acquired information. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 357(1427):1559–66, nov 2002. ISSN 0962-8436. doi: 10.1098/rstb.2002.1065. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1693065{&}tool=pmcentrez{&}rendertype=abstract`.

S. Goss, J. Aron, J.-L. Deneubourg, and J. M. Pasteels. Self-organized Shortcuts in the Argentine Ant. *Naturwissenschaften*, 76:579–581, 1989.

C. Grüter and E. Leadbeater. Insights from insects about adaptive social information use. *Trends in Ecology & Evolution*, pages 1–8, feb 2014. ISSN 01695347. doi: 10.1016/j.tree.2014.01.004. URL `http://linkinghub.elsevier.com/retrieve/pii/S0169534714000172`.

R. Hertwig and I. Erev. The description-experience gap in risky choice. *Trends in Cognitive Sciences*, 13(12):517–523, 2009. ISSN 13646613. doi: 10.1016/j.tics.2009.09.004.

W. Hoppitt and K. N. Laland. *Social Learning: An Introduction to Mechanisms, Methods, and Models*. Princeton University Press, 2013. ISBN 0691150702, 9780691150703.

E. Jouini, C. Napp, and D. Nocetti. Collective risk aversion. *Social Choice and Welfare*, 40(2):411–437, 2013. ISSN 01761714. doi: 10.1007/s00355-011-0611-9.

T. Kameda and D. Nakanishi. Cost–benefit analysis of social/cultural learning in a nonstationary uncertain environment. *Evolution and Human Behavior*, 23(5):373–393, 2002. ISSN 10905138. doi: 10.1016/S1090-5138(02)00101-0. URL `http://www.sciencedirect.com/science/article/pii/S1090513802001010`.

A. Kandler and K. N. Laland. Tradeoffs between the strength of conformity and number of conformists in variable environments. *Journal of theoretical biology*, 332:191–202, sep 2013. ISSN 1095-8541. doi: 10.1016/j.jtbi.2013.04.023. URL `http://www.ncbi.nlm.nih.gov/pubmed/23643630`.

R. L. Kendal, I. Coolen, Y. van Bergen, and K. N. Laland. Trade-Offs in the Adaptive Use of Social and Asocial Learning. *Advances in the Study of Behavior*, 35(05):333–379, 2005. ISSN 00653454. doi: 10.1016/S0065-3454(05)35008-X.

A. J. King and G. Cowlishaw. When to use social information: the advantage of large group size in individual decision making. *Biology letters*, 3(2):137–139, 2007. ISSN 1744-9561. doi: 10.1098/rsbl.2007.0017.

E. A. Ludvig, C. R. Madan, J. M. Pisklak, and M. L. Spetch. Reward context determines risky choice in pigeons and humans. *Biology Letters*, 10(8), 2014. ISSN 1744957X. doi: 10.1098/rsbl.2014.0451.

J. March. Learning to be risk averse. *Psychological Review*, 103(2):309–319, 1996. ISSN 1939-1471. doi: 10.1037/0033-295X.103.2.309.

R. McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.

R. McElreath, M. Lubell, P. J. Richerson, T. M. Waring, W. Baum, E. Edsten, C. Efferson, and B. Paciotti. Applying evolutionary models to the laboratory study of social learning. *Evolution and Human Behavior*, 26(6):483–508, nov 2005. ISSN 10905138. doi: 10.1016/j.evolhumbehav.2005.04.003. URL `http://linkinghub.elsevier.com/retrieve/pii/S1090513805000231`.

R. McElreath, A. V. Bell, C. Efferson, M. Lubell, P. J. Richerson, and T. Waring. Beyond existence and aiming outside the laboratory: estimating frequency-dependent and pay-off-biased social learning strategies. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1509):3515–28, nov 2008. ISSN 1471-2970. doi: 10.1098/rstb.2008.0131. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2607339{&}tool=pmcentrez{&}rendertype=abstract`.

E. Miu, N. Gulley, K. N. Laland, and L. Rendell. Innovation and cumulative culture through tweaks and leaps in online programming contests. *Nature Communications*, 9(1):1–8, 2018. ISSN 20411723. doi: 10.1038/s41467-018-04494-0.

T. J. H. Morgan, N. T. Uomini, L. E. Rendell, L. Chouinard-Thuly, S. E. Street, H. M. Lewis, C. P. Cross, C. Evans, R. Kearney, I. de la Torre, a. Whiten, and K. N. Laland. Experimental evidence for the co-evolution of hominin tool-making teaching and language. *Nature Communications*, 6:6029, jan 2015. ISSN 2041-1723. doi: 10.1038/ncomms7029. URL http://www.nature.com/doifinder/10.1038/ncomms7029.

M. Moussaïd, H. Brighton, and W. Gaissmaier. The amplification of risk in experimental diffusion chains. *Proceedings of the National Academy of Sciences*, page 201421883, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1421883112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1421883112.

M. Muthukrishna, T. J. Morgan, and J. Henrich. The when and who of social learning and conformist transmission. *Evolution and Human Behavior*, 2015. ISSN 10905138. doi: 10.1016/j.evolhumbehav.2015.05.004. URL http://linkinghub.elsevier.com/retrieve/pii/S1090513815000586.

A. Najar, E. Bonnet, B. Bahrami, and S. Palminteri. The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLoS biology*, pages 1–25, 2020. doi: 10.1371/journal.pbio.3001028. URL http://dx.doi.org/10.1371/journal.pbio.3001028.

W. Nakahashi, J. Y. Wakano, and J. Henrich. Adaptive social learning strategies in temporally and spatially varying environments : how temporal vs. spatial variation, number of cultural traits, and costs of learning influence the evolution of conformist-biased transmission, payoff-biased transmissi. *Human nature (Hawthorne, N.Y.)*, 23(4): 386–418, dec 2012. ISSN 1936-4776. doi: 10.1007/s12110-012-9151-y. URL http://www.ncbi.nlm.nih.gov/pubmed/22926986.

R. M. Raafat, N. Chater, and C. Frith. Herding in humans. *Trends in cognitive sciences*, 13(10):420–8, oct 2009. ISSN 1364-6613. doi: 10.1016/j.tics.2009.08.002. URL http://www.ncbi.nlm.nih.gov/pubmed/19748818.

L. Real and T. Caraco. Risk and foraging in stochastic environments. *Annual review of ecology and systematics*, 17 (1):371–390, 1986.

L. Real, J. Ott, and E. Silverfine. On the tradeoff between the mean and the variance in foraging: effect of spatial distribution and color preference. *Ecology*, 63(6):1617–1623, 1982.

L. A. Real. Uncertainty and Pollinator-Plant Interactions: The Foraging Behavior of Bees and Wasps on Artificial Flowers. *Ecology*, 62(1):20–26, 1981. ISSN 0012-9658. doi: 10.2307/1936663.

L. Rendell, R. Boyd, D. Cownden, M. Enquist, K. Eriksson, M. W. Feldman, L. Fogarty, S. Ghirlanda, T. Lillicrap, and K. N. Laland. Why copy others? Insights from the social learning strategies tourna-

ment. *Science (New York, N.Y.)*, 328(5975):208–13, apr 2010. ISSN 1095-9203. doi: 10.1126/science.1184719. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2989663{&}tool=pmcentrez{&}rendertype=abstract.

T. Sasaki and D. Biro. Cumulative culture can emerge from collective intelligence in animal groups. *Nature Communications*, 8:1–6, 2017. ISSN 20411723. doi: 10.1038/ncomms15049. URL http://dx.doi.org/10.1038/ncomms15049.

T. Seeley, S. Camazine, and J. Sneyd. Collective decision-making in honey bees: how colonies choose among nectar sources. *Behavioral Ecology and Sociobiology*, 28:277–290, 1991. URL http://www.springerlink.com/index/J352RUP1J10W5082.pdf.

R. S. Shupp and A. W. Williams. Risk preference differentials of small groups and individuals. *Economic Journal*, 118(525):258–283, 2008. ISSN 00130133. doi: 10.1111/j.1468-0297.2007.02112.x.

D. J. T. Sumpter. The principles of collective animal behaviour. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 361(1465):5–22, jan 2006. ISSN 0962-8436. doi: 10.1098/rstb.2005.1733. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1626537{&}tool=pmcentrez{&}rendertype=abstract.

T. Sumpter and C. Pratt. A modelling framework for understanding social insect foraging. *Behavioral Ecology and Sociobiology*, 53(3):131–144, 2003. doi: 10.1007/s00265-002-0549-0.

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

S. Suzuki, E. L. Jensen, P. Bossaerts, and J. P. O'Doherty. Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. *Proceedings of the National Academy of Sciences of the United States of America*, 113(14):3755–3760, 2016. ISSN 10916490. doi: 10.1073/pnas.1600092113.

W. Toyokawa, H.-R. Kim, and T. Kameda. Human Collective Intelligence under Dual Exploration-Exploitation Dilemmas. *PloS one*, 9(4):e95789, jan 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0095789. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3995913{&}tool=pmcentrez{&}rendertype=abstract.

W. Toyokawa, Y. Saito, and T. Kameda. Individual differences in learning behaviours in humans: asocial exploration tendency does not predict reliance on social learning. *Evolution and Human Behavior*, 38(3):325–333, 2017. ISSN

10905138. doi: 10.1016/j.evolhumbehav.2016.11.001. URL http://linkinghub.elsevier.com/retrieve/pii/S1090513816301180.

W. Toyokawa, A. Whalen, and K. N. Laland. Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, page 326637, 2019. ISSN 2397-3374. doi: 10.1101/326637. URL https://www.biorxiv.org/content/early/2018/06/19/326637.

P. Ward and A. Zahavi. The Importance of Certain Assemblages of Birds As Information Centres for Food-Finding. *Ibis*, 115(4):517–534, 1973. ISSN 1474919X. doi: 10.1111/j.1474-919X.1973.tb01990.x.

E. U. Weber. Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). *Climatic Change*, 77(1-2):103–120, 2006. ISSN 01650009. doi: 10.1007/s10584-006-9060-3.

E. U. Weber, S. Shafir, and A. R. Blais. Predicting Risk Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation. *Psychological Review*, 111(2):430–445, 2004. ISSN 0033295X. doi: 10.1037/0033-295X.111.2.430.

E. Yechiam and J. R. Busemeyer. The effect of foregone payoffs on underweighting small probability events. *Journal of Behavioral Decision Making*, 19(1):1–16, 2006. ISSN 10990771. doi: 10.1002/bdm.509.

E. Yechiam, I. Erev, and G. Barron. The effect of experience on using a safety device. *Safety Science*, 44(6):515–522, 2006. ISSN 09257535. doi: 10.1016/j.ssci.2005.11.006.

7 Acknowledgements

This work was supported by the Small Project Grant from the Centre for the Advanced Study of Collective Behaviour, the university of Konstanz (S20-06), and by University of Konstanz Committee on Research (FP031/19). We thank Lucy Aplin and Brendan Barrett for many helpful comments on earlier versions of this paper.

8 Author contributions

W.T planned the study, built and analysed the mathematical model, made the experimental material, and run the online experiment. W.T and W.G analysed the data, discussed and refined the experimental design, and wrote the manuscript.

9 Materials & Correspondence

Correspondence and requests for materials should be addressed to W.T.

10 Supplementary materials

10.1 Supplementary methods

10.1.1 An analytical result derived by Denrell (2007)

In the simplest setup of the two-armed bandit task, Denrell (2007) derived an explicit form for the asymptotic probability of choosing the risky alternative $P_r^\star$ (as $t \to \infty$) as follows:

$$P_r^\star = \frac{1}{1 + \exp\left[\frac{\alpha\beta^2 \text{s.d.}^2}{2(2-\alpha)} - \beta(\mu - \pi_s)\right]}. \tag{4}$$

The equation 4 identifies a condition under which reinforcement learners exhibit risk aversion. Actually, when there is no risk premium (i.e. $\mu \leq \pi_s$), the condition of risk aversion always holds, i.e. $P_r^\star < 0.5$. Consider the case where risk aversion is suboptimal, that is $\mu > \pi_s$. Eq. 4 suggests that the suboptimal risk aversion emerges when learning is myopic (i.e. α is large) and/or decision-making is less explorative (i.e. β is large). For instance, when the payoff distribution of the risky alternative is set $\mu = \pi_s + 0.5$ and s.d.$^2 = 1$, the condition of risk-aversion, $P_r^\star < 0.5$, holds under $\beta > (2-\alpha)/\alpha$ which corresponds to the area above the dashed curve in Figure 1. Risk aversion becomes further prominent when the risk premium $\mu - \pi_s$ is small and/or the payoff variance s.d.^{2} is large.

10.1.2 The online experiments

Participants. A total of 492 subjects who declared their residence in the UK, the US, Ireland, or Australia (234 self-identified as women, 211 as men, and 1 as others; mean (minimum, maximum) age = 34.8 (18, 74) years) participated in the task through Prolific Academic. All participants consented to participation through an online consent form at the beginning of the task. We excluded subjects who disconnected from the online task before completing at least the first 35 rounds from our computational model-fitting analysis, resulting in 467 subjects (the detailed distribution of subjects for each condition is shown in Table 1). The task was only available for English-speaking subjects and they must be 18 years old or older. Only participants who passed a comprehension quiz at the end of the instruction could enter the task. Participants were paid 0.8 GBP for a show-up fee as well as an additional bonus payment depending on their performance in the decision-making task (500 artificial points were converted to 8 pence, resulting in a bonus ranging between £ 1.0 and £ 3.5).

Sample size. Our target sample size was 50 participants for each of the individual condition and 150 participants for each of the group condition. For the group condition, our target average group size was 5 individuals per group.

Participants each played the task for 70 trials. This trial number had been justified by a model recovery analysis of a previous study (Toyokawa et al., 2019).

Due to the nature of the 'waiting lobby' available only for 3 minutes, we could not fully control the exact size of each experimental group. Therefore, we set the maximum capacity of a lobby to 8 individuals for the 1-safe-1-risky task which was conducted in August 2020, so as to buffer potential drop-outs during a waiting period. Since we learnt that drop-out happened rather rarely than we originally expected (Supporting Fig. 16), we reduced the lobby capacity to 6 for both the 1-risky-3-safe and 2-risky-2-safe tasks, which were conducted in October 2020. As a result, we had 20 groups (mean group size = 6.95), 21 groups (mean group size = 4.7), and 19 groups (mean group size = 4.3), for the 1-risky-1-safe, 1-risky-3-safe, and 2-risky-2-safe task, respectively. Although we could not achieve the sample size targeted, partly due to the drop-outs during the task and to a fatal error occurred in the experimental server in the first few sessions of the 4-armed tasks, the parameter recovery test with $N = 105$ suggested that the current sample size should be reliable enough to estimate social influences for each participant (Supporting Fig. 15).

10.1.3 The hierarchical Bayesian parameter estimation

We used hierarchical Bayesian method (HBM) to estimate the free parameters of our learning model. HBM allows us to estimate individual differences, while this individual variation is bounded by the group-level (i.e. hyper) parameters.

We assumed the Student's t distributions for individual random effects of each parameter so as to allow a few 'outliers', because the Student's t distribution has a longer tail compared to a normal distribution. To do so, we used the following non-centred reparameterization (the 'Matt trick') as follows:

$$\text{logit}(\alpha_i) = \mu_\alpha + \upsilon_\alpha * \alpha_{\text{raw},i}$$

where μ_α is a global mean of logit(α_i) and υ_α is a global scale parameter of the individual variations, which is multiplied by a standardised individual random variable $\alpha_{\text{raw},i}$. We used a standardised normal prior distribution for μ_α and an exponential prior for υ_α. The same method was applied to other learning parameters β_i, σ_i, and θ_i.

We assumed that the 'raw' values of individual random variables ($\alpha_{\text{raw},i}$, $\beta_{\text{raw},i}$, $\sigma_{\text{raw},i}$, $\theta_{\text{raw},i}$) were drawn from a multivariate normal distribution. The correlation matrix was estimated using a Cholesky decomposition with a weakly informative Lewandowski-Kurowicka-Joe (LKJ) prior that was sceptical about very high or very low correlations between the parameters (McElreath, 2020).

Model fitting. All models were fitted using the Hamiltonian Monte Carlo engine CmdStan 2.25.0 (`https://mc-stan.org/cmdstanr/index.html`) in R 4.0.2 (`https://www.r-project.org`). The models contained at least six parallel chains and we confirmed convergence of the MCMC using both the Gelman–Rubin statistics satisfied $\hat{R} \leq 1.01$ and the effective sample sizes greater than 500. To improve convergence, we employed the non-centred reparameterization coupled with the Cholesky decomposition of correlation effects (McElreath, 2020; Deffner et al., 2020). The R and Stan code used in the model fitting are available from an online repository.

10.1.4 The value-shaping social influence model

We considered another implementation of social influences in the reinforcement learning, namely, the value-shaping process (Biele et al., 2011; Najar et al., 2020) rather than the decision-biasing process assumed in our main analyses. In the value-shaping model, social influence modifies the Q value's updating process as follows:

$$Q_{i,t+1} \leftarrow (1-\alpha)Q_{i,t} + \alpha(\pi_{i,t} + \sigma_{vs}\bar{\pi}\frac{N^{\theta}_{i,t-1}}{N^{\theta}_{s,t-1} + N^{\theta}_{r,t-1}})$$

where the social frequency cue acts as an additional 'bonus' of the value that was weighted by σ_{vs} ($\sigma_{vs} > 0$) and standardised by the expected payoff from choosing randomly among all alternatives $\bar{\pi}$. Here we assumed no direct social influence on the action selection process (i.e. $\sigma = 0$ in our main model). We confirmed that the collective behavioural rescue could emerged when the inverse temperature β is sufficiently small (Supporting Fig. 10). Although it is out of the focus of this article, whether people's behaviour can be fit better by this value-shaping model than by the decision-biasing model is an interesting question for the future research (Najar et al., 2020).

10.2 Supplementary figures and tables

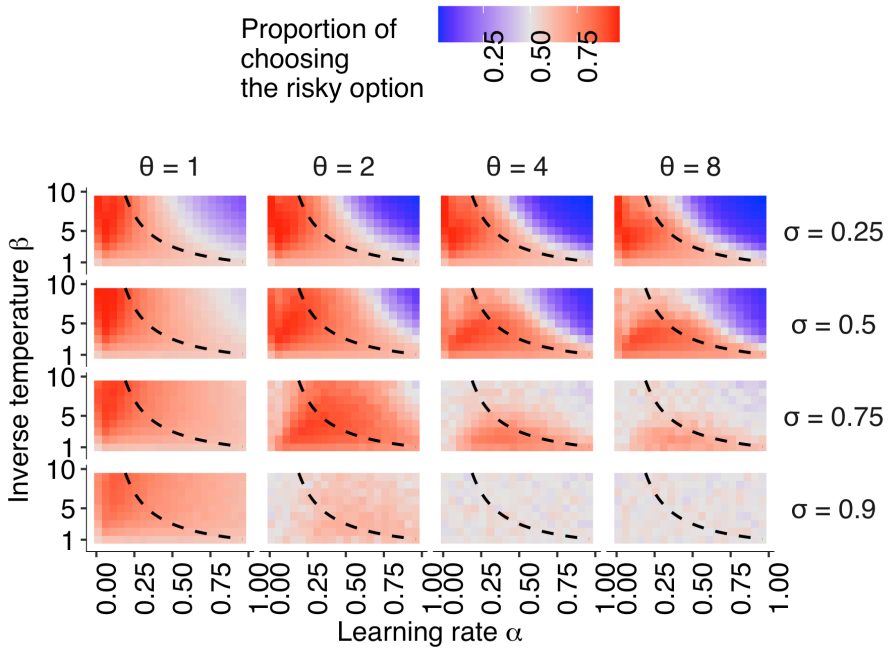

Figure 7: The relationships between individual learning rate (α) and individual inverse temperature (β) across the different combinations of social learning parameters on the mean proportion of choosing the risky alternative in the second half of the trials of the two-armed bandit task described in Fig.1. The dashed curve gives a set of parameter combinations with which asocial learners are expected to choose the risky alternative as same proportion as they choose the safe alternative (i.e. $P_r^\star = 0.5$) in the infinite time horizon $T \to \infty$, given by $\beta = (2-\alpha)/\alpha$.

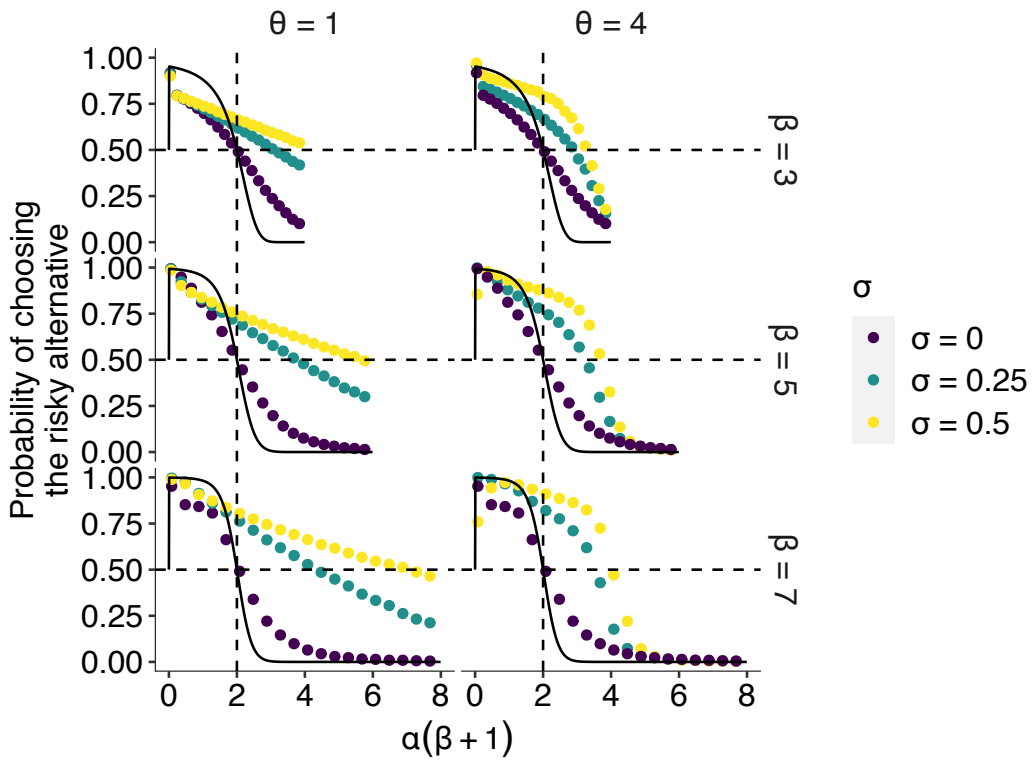

Figure 8: The effect of social learning on the average decision performance in the longer time horizon. The x-axis is an interaction of two reinforcement learning parameters $\alpha(\beta+1)$, that is, the susceptibility to hot stove effect. The y-axis is the mean proportion of choosing the optimal risky alternative in the last 75 trials in the two-armed bandit task whose setup was the same as in Fig. 1 and 2 (i.e. $\mu = 0.5$, $sd = 1$) except for the longer time horizon $T = 1075$ compared to the time horizon used in the main text ($T = 150$). The dotted curves are the mean result of agent-based simulations of groups of social learners with two different mean values of the copying weight $\sigma \in \{0.25, 0.5\}$ or individual learners with $\sigma = 0$. Each panel shows different combinations of the inverse temperature β and the conformity exponent θ. The black solid curve is the theoretical benchmark where individual reinforcement learners were expected to asymptote with $T \to +\infty$. Compared to Fig. 2 in the main text, individual learners became closer to the benchmark. On the other hand, the performance of social learners could remain deviated from the benchmark, suggesting that social influence had a qualitative impact on the course of learning and decision-making, rather than merely slowing-down approaching to the equilibrium of individual learning.

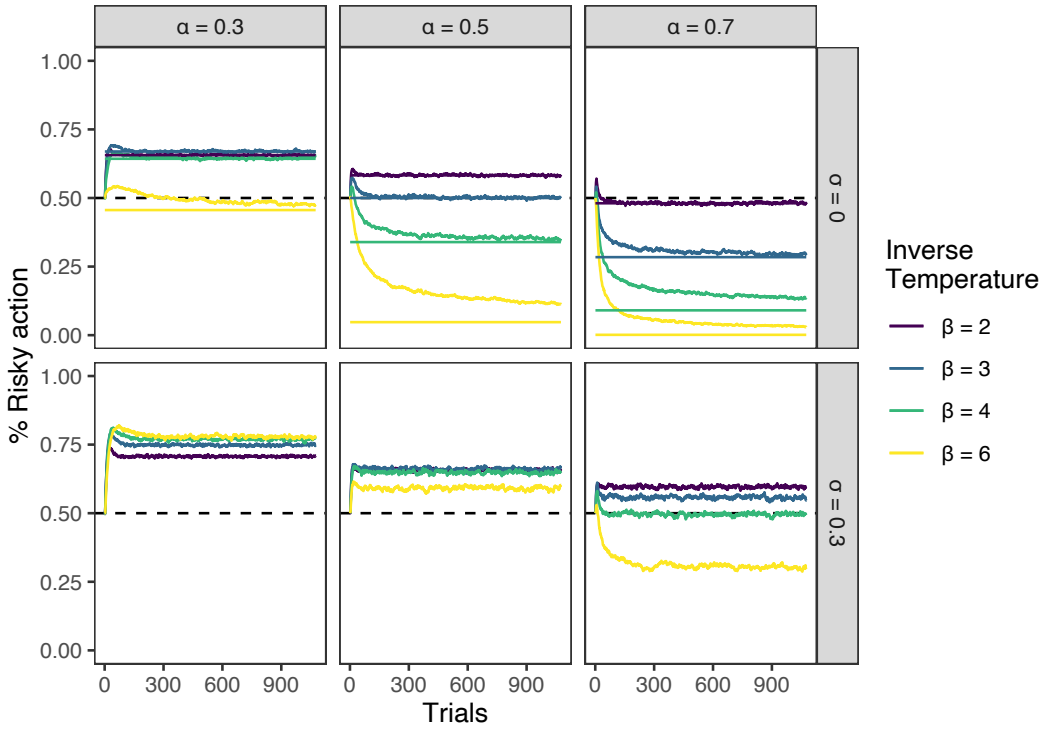

Figure 9: The effect of social learning on the time evolution of decision performance. The x-axis is the number of trial. The y-axis is the mean proportion of choosing the optimal risky alternative. Each colour shows different β. For the asocial learning condition (that is, $\sigma = 0$), the analytical benchmark to which reinforcement learners asymptote is shown as a horizontal lines. Conformity exponent θ is 2. Group size is 8. The simulation was repeated 1000 times for each combination of parameters. Comparing to the asocial learning cases, the social learning ($\sigma = 0.3$) qualitatively alters the course of learning, rather than just speeding-up or slowing-down learning.

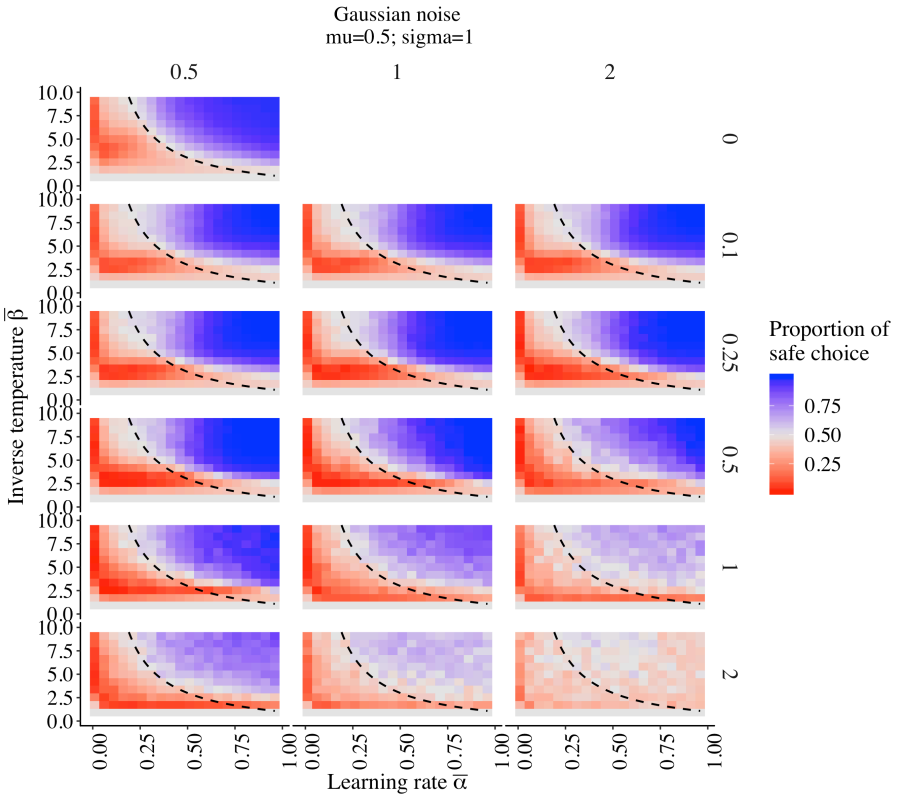

Figure 10: The result of the value-shaping social influence model. The relationships between individual learning rate (α) and individual inverse temperature (β) across the different social learning parameters' combination on a degree of the risk aversion. Different social learning weights (σ_{vs}) are shown from the topmost to bottommost ($\sigma_{vs} \in \{0, 0.1, 0.25, 0.5, 1, 2\}$). Different conformity exponents are shown from the leftmost to rightmost ($\theta \in \{0.5, 1, 2\}$). The dashed curve is the asymptotic equilibrium at which asocial learners are expected to end up choosing both alternatives equally likely (i.e. $P_r^\star = 0.5$), given by $\beta = (2 - \alpha)/\alpha$.

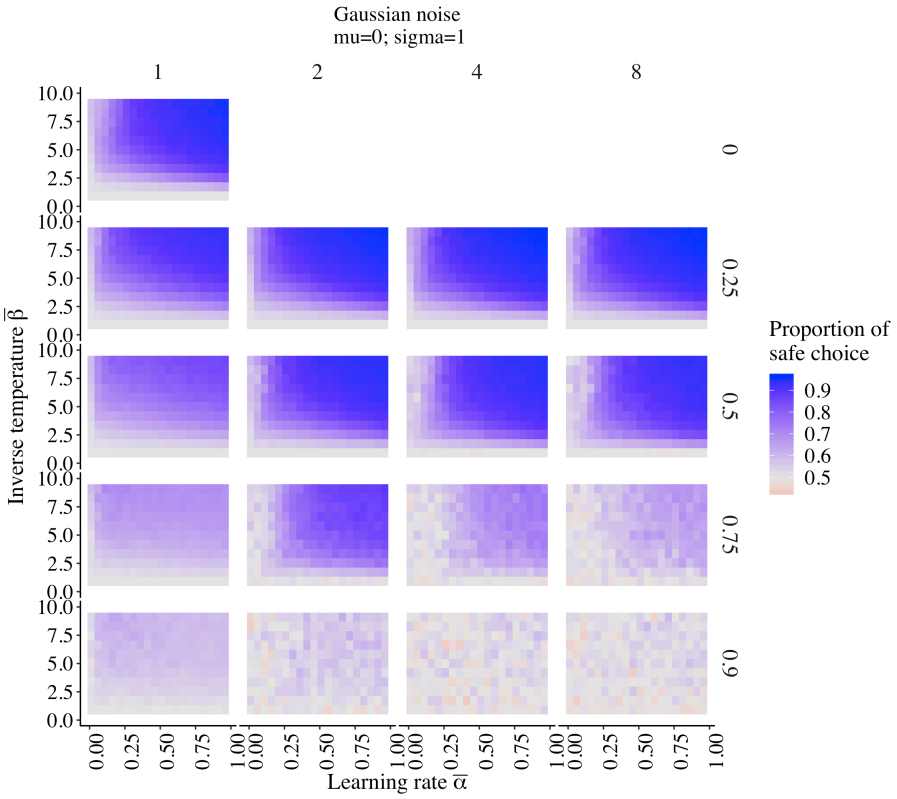

Figure 11: The relationships between individual learning rate (α) and individual inverse temperature (β) across the different social learning parameters' combination on a degree of the risk aversion. Different social learning weights (σ) are shown from the topmost to bottommost ($\sigma \in \{0, 0.25, 0.5, 0.75, 0.9\}$). Different conformity exponents are shown from the leftmost to rightmost ($\theta \in \{1, 2, 4, 8\}$). The risk premium is negative $\mu = -0.5$.

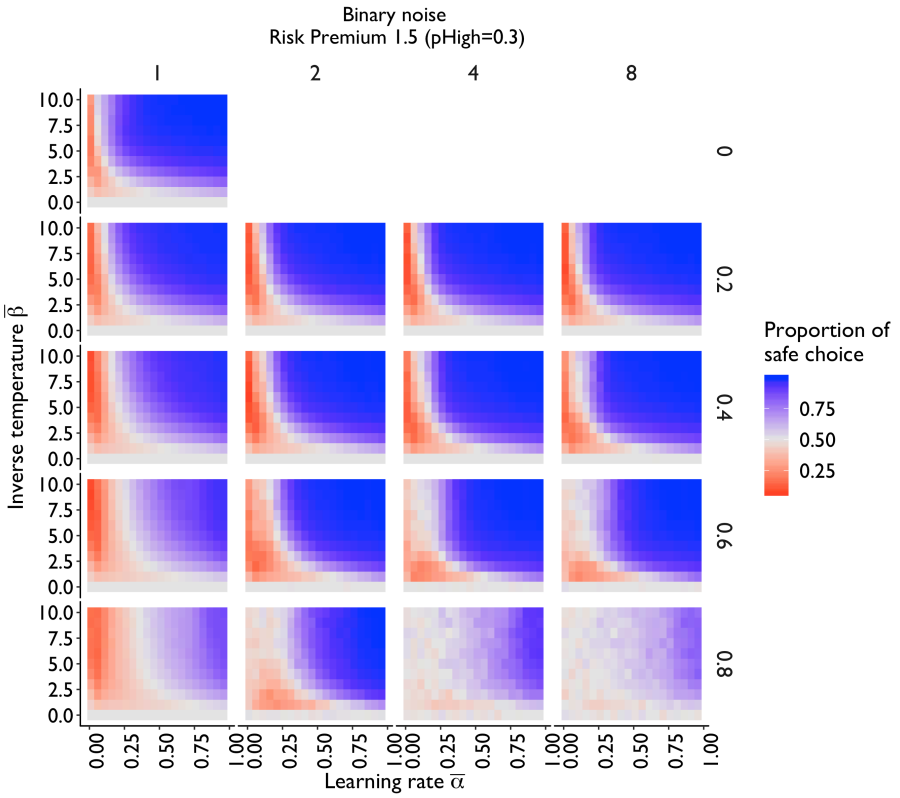

Figure 12: The relationships between individual learning rate (α) and individual inverse temperature (β) across the different social learning parameters' combination on a degree of the risk aversion. Different social learning weights (σ) are shown from the topmost to bottommost ($\sigma \in \{0, 0.25, 0.5, 0.75, 0.9\}$). Different conformity exponents are shown from the leftmost to rightmost ($\theta \in \{1, 2, 4, 8\}$). The binary payoff distribution was used where the safe alternative always gives $\pi_s = 1$ while the risky alternative gives wither $\pi_r = 0$ at 70% of chance and $\pi_r = 5$ at 30% of chance. The risk premium was 1.5.

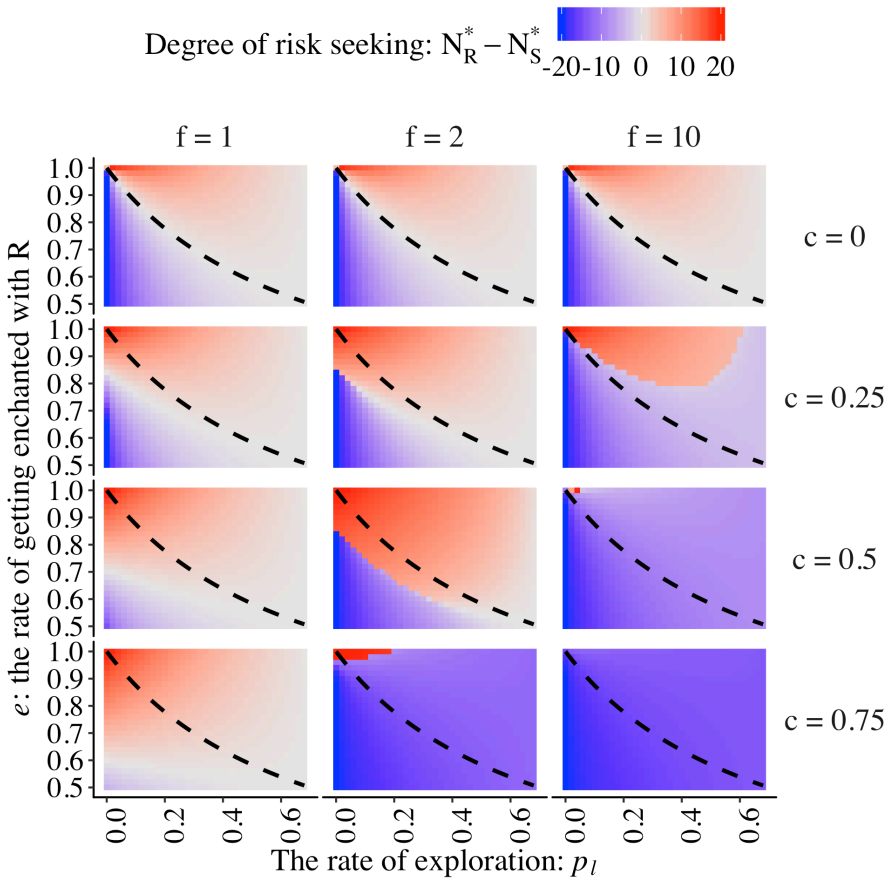

Figure 13: The effect of both the per individual probability of exploration p_l and e (i.e. the ratio of individuals who prefer behavioural state R) on the equilibrium degree of risk-seeking (i.e. $N_R^* - N_S^*$), across the different combinations of social influence parameters. Different social influence weights are shown from the top to bottom panels ($\sigma \in \{0, 0.25, 0.5, 0.75\}$). Different conformity exponents are shown from the left to right panels ($\theta \in \{1, 2\}$). The dashed curve is $e = p_h/(p_h + p_l)$. The numeric solution was obtained with conditions $N_{S,t=0}^- = N_{S,t=0}^+ = 5$, $N_{R,t=0} = 10$, and $p_h = 0.7$.

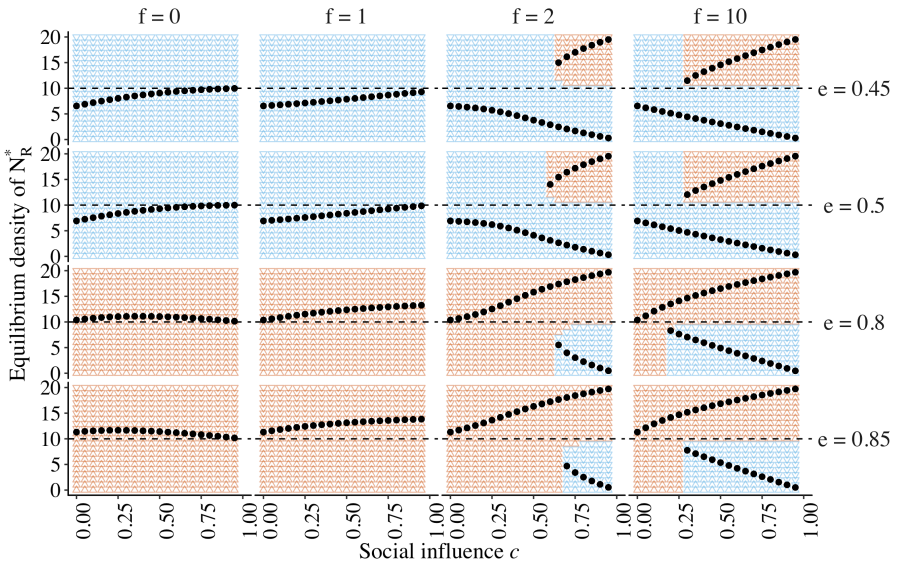

Figure 14: The relationships between the social influence weight c and the equilibrium number of the individuals choosing the risky alternative $N_R^\star$ across the different conformity exponent $\theta(\in \{0, 1, 2, 10\})$, shown in the black dots. The triangular points shown in the background of each panel indicates regions in which the group approaches to the risk aversion (i.e. $N_R^\star < 10$; blue) or to the risk seeking equilibrium (i.e. $N_R^\star > 10$; red). Two different equilibria mean that the system has a bifurcation under a given σ. The shape of the background triangles indicates whether the N_R increases ($\triangle$) or decreases ($\bigtriangledown$) relative to its starting position. The other parameters are set to $p_h = 0.7$, $p_l = 0.2$, $e = 0.7$.

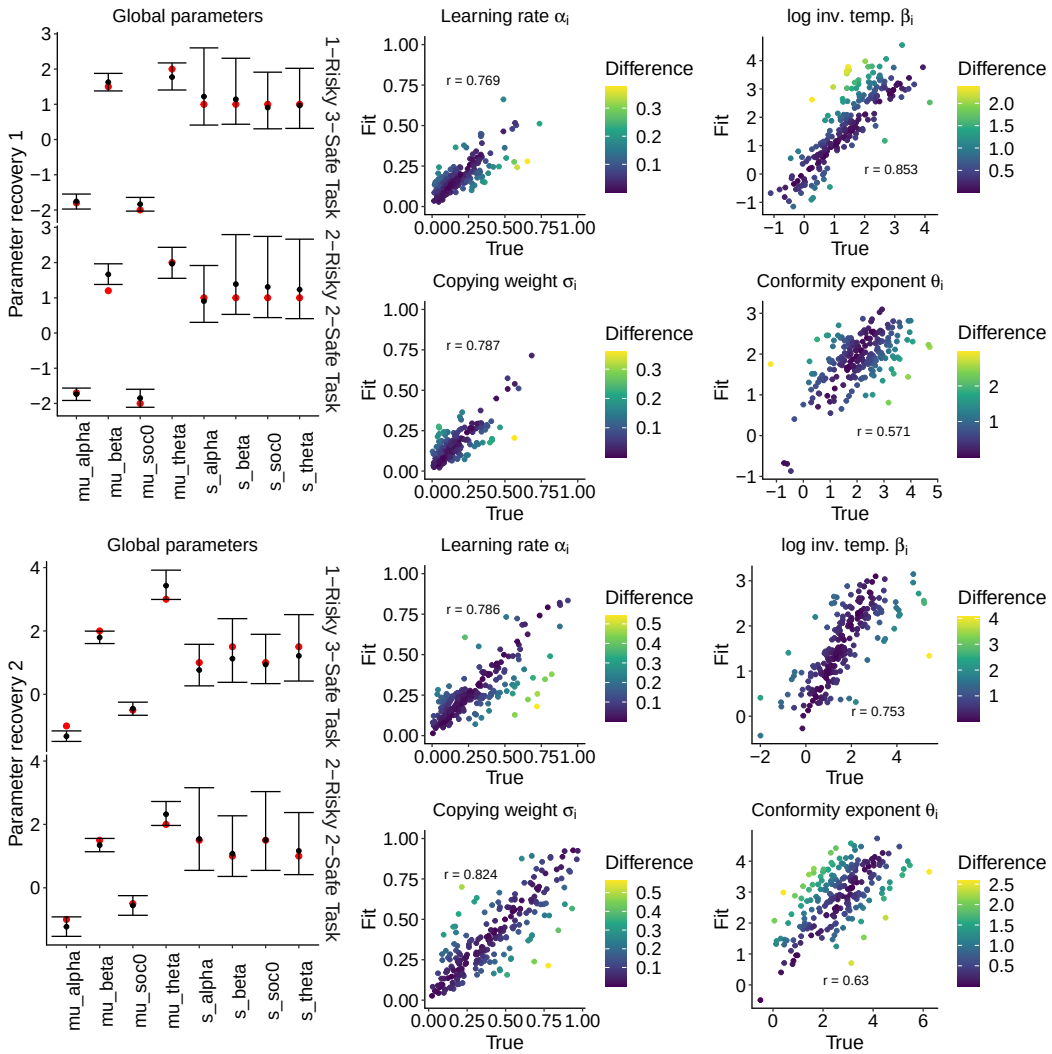

Figure 15: The parameter recovery performance for (left column) the global parameters fitted for each of the two different 4-armed bandit tasks, that are the 1-risky 2-safe task ($N = 105$) and the 2-risky 2-safe task ($N = 105$), and for (right column) individual-level parameters shown across the two task conditions ($N = 210$). (Left column) The red points are the true values and the black points are the mean posterior values (i.e. recovered values). The 95% Bayesian credible intervals are shown by the error bars. (Right column) The x-axis is the true value and the y-axis is the fitted (i.e. the mean posterior) individual value. The differences between the true value and the estimated value are shown in different colours (Dark: fit well). The Pearson's correlation coefficients between the true and fitted values are shown. The top row and bottom row of the panels are the results of parameter recovery test 1 and 2, respectively.

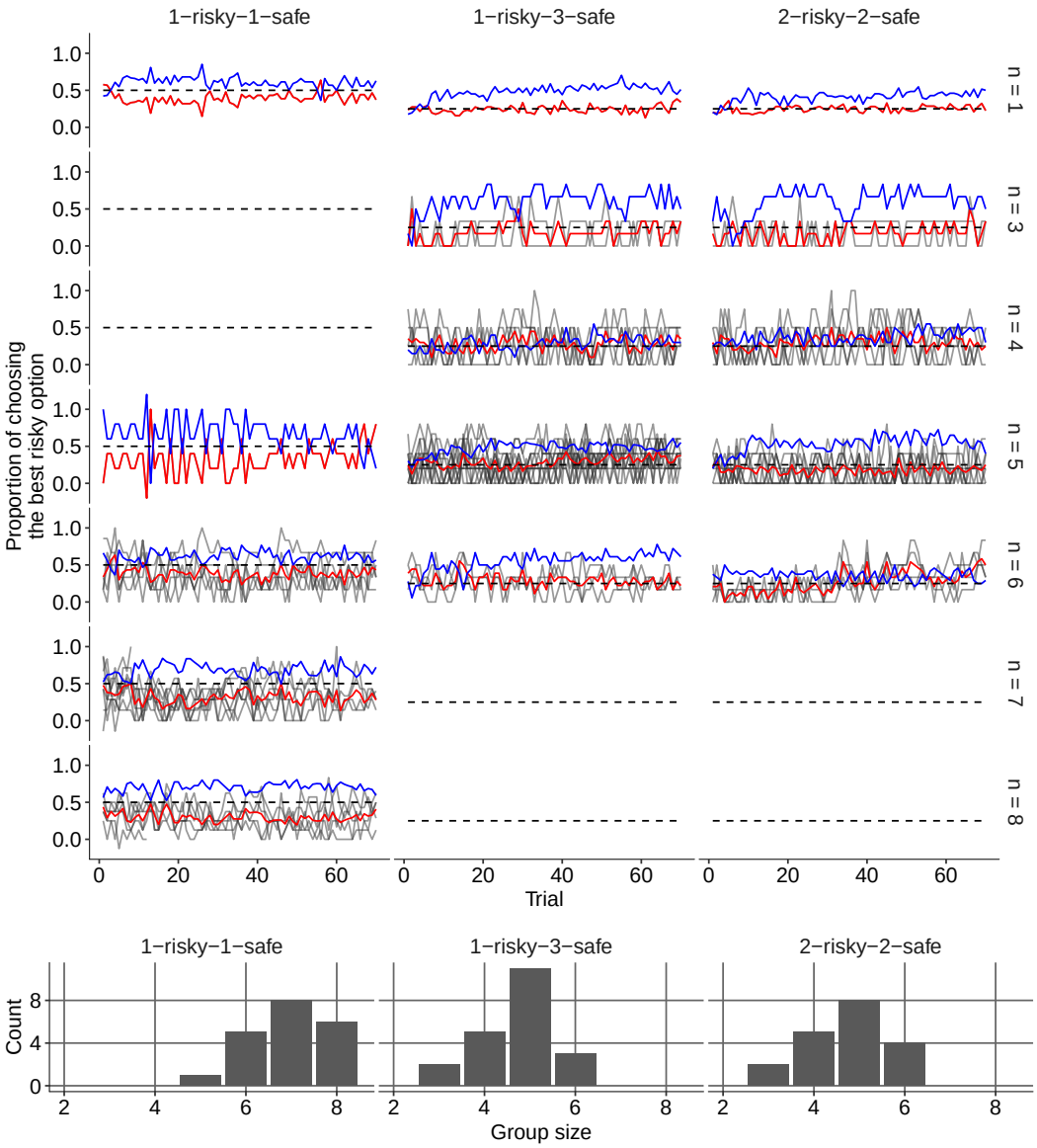

Figure 16: Dynamics of the experimental groups' mean performance and the distribution of group size. Group size is indicated in the righthand side. Each group's mean proportion of choosing the optimal risky option is shown in grey, whose mean value across the same group size category is shown in red. The blue line is the mean proportion of choosing the best safe option (that is suboptimal) averaged across the same group size category.