Social learning with a grain of salt

Alexandra Witt[1] **(alexandra.witt@gmx.net), Wataru Toyokawa**[2]**, Kevin Lala**[3,§]**, Wolfgang Gaissmaier**[2] **& Charley M. Wu**[1]

[1] Human and Machine Cognition Lab, University of Tübingen, Tübingen, Germany
[2] University of Konstanz, Konstanz, Germany
[3] University of St Andrews, St Andrews, UK
[§] formerly Laland

Abstract

Humans are remarkably effective social learners, with several recent studies formalizing this capacity using computational models. However, previous research has often been limited to tasks where observer and demonstrator share the same reward function. In contrast, humans can learn from others who have different preferences, skills, or goals. To study social learning under individual differences, we introduce the *socially correlated bandit*, where participants have personalized rewards, which are correlated with but not identical to those of others. Social information can still be useful, but not when used verbatim. We present a model of *Social Generalization* that integrates individual and social information into the generalization process, but assumes social information to be noisier and thus less informative. This model out-competes previous models, with it being the dominant strategy in evolutionary simulations. Our findings expand on previous models of social learning, showing humans can integrate social information more flexibly than previously assumed.

Keywords: social learning; computational modelling; collective intelligence; generalization; Gaussian process

Introduction

Social learning is an extremely adaptive ability, to which we humans owe much of our evolutionary success (Heyes, 2018; Henrich, 2015; Laland, 2017). After all, being able to learn from the experiences of others not only saves time and effort, but may even help avoid disappointing or even harmful outcomes (Laland, 2004). A simple example would be to rely on average restaurant ratings when looking for a place to eat in a new city, instead of randomly selecting an option you have no information about.

However, social information cannot always be used verbatim. The subjective value of an option depends on many factors, such as one's individual circumstances, goals, skills, and preferences (Wu, Vélez, & Cushman, 2022; FeldmanHall & Nassar, 2021). In the restaurant example, while higher quality restaurants will likely have higher ratings, other reviewers can have very different subjective preferences, such as a different degree of spice tolerance. Thus, it would be a mistake to use social information in the same fashion as one's own experience. In fact, it seems far more common that social observations would need to be taken with a grain of salt, compared to cases where social information can be simply used verbatim, based on an objective and universal reward function.

Yet much of the previous literature has focused on settings where both demonstrator and observer share the exact same task environment and outcomes (Charpentier, Iigaya, & O'Doherty, 2020; Toyokawa, Whalen, & Laland, 2019; Najar, Bonnet, Bahrami, & Palminteri, 2020; Naito, Katahira, & Kameda, 2022). This work has uncovered evidence that people flexibly integrate social information with individual learning through exact imitation (Toyokawa et al., 2019) or by enhancing the subjective value of the choices made by a demonstrator (Najar et al., 2020). However, we cannot be sure that the same mechanisms equally apply in settings where outcomes may differ on an individual basis, such as when reading restaurant or product reviews, or in other matters of taste.

While there is some work on how to integrate social information in matters of taste (Analytis, Barkoczi, & Herzog, 2018; Müller-Trede, Choshen-Hillel, Barneron, & Yaniv, 2018; Yaniv, Choshen-Hillel, & Milyavsky, 2011), this line of research is largely theoretical and has not proposed any computational models of how humans behave in such situations. Thus, how we use social information effectively despite our differences remains an unanswered question.

Goals and Scope

To investigate social learning in contexts with unique but correlated rewards across individuals, we introduce the socially correlated bandit. In this paradigm, the highest rewards are generally located in the same region for all participants, but directly copying another participant's choices will not usually lead to the maximum payoff. This mimics real-world settings where social information is colored by individual differences in preferences and circumstances: while some standards apply to everyone, the option someone else values most highly will not necessarily be the most rewarding to you.

Participants explored these socially correlated environments in groups of four, with full information about other participants' choices and reward outcomes. We used evolutionary simulations across multiple social learning models to find the normatively best strategy, which was our novel Social Generalization (SG) model. We then fit these models to the behavioural data collected in the experiment and found that participants' behaviour was also most accurately predicted by SG. This shows that humans are able to integrate social information with more nuance than assumed by models from previous literature.

In the following, we first introduce the task in more detail, as well as a set of candidate reinforcement learning models integrating social information at different stages of the

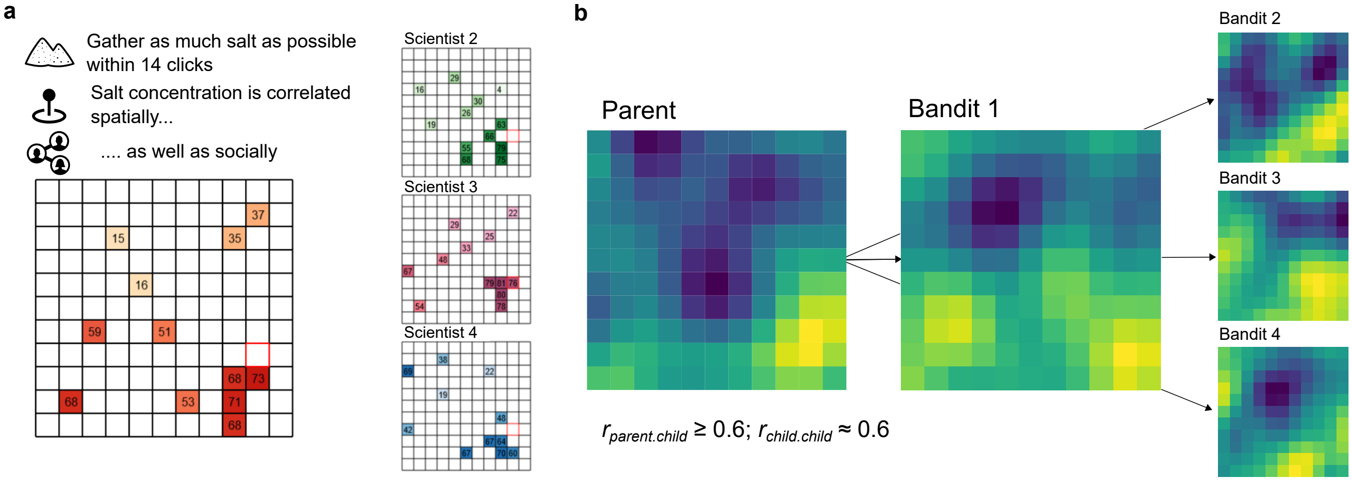

Figure 1: Task. **a**) Socially correlated bandit task. Participants searched for rewards in groups of four and got updates on choices and outcomes of other group members. The cover story described a goal of gathering salt samples from alien oceans in a team of scientists. Each scientist is seeking a different type of salt, which are correlated due to similar generation processes underwater. **b**) For each set of reward environments, we first generated a parent grid with spatially correlated rewards. The parent environment was then used to generate four children bandits, which were correlated with one another.

decision-making process. We then explain the evolutionary simulations we used to determine which populations of models evolve from a variety of starting populations. Lastly, we analyze and fit models to behavioral data from groups of human participants.

Methods

The socially correlated bandit is an extension of past work using a spatially correlated bandit (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018), where tiles on a grid represent the different arms of the bandit. Each tile generates noisy rewards, but reward expectations are distributed with a smoothly varying spatial structure (i.e., nearby tiles tend to have similar rewards). Participants are asked to maximize rewards under a limited search horizon, meaning they cannot sample all possibilities to find the optimum. However, the spatial correlations allow participants to infer the expected value of unobserved arms by generalizing from a limited set of past observations.

Here, we add social correlations, with groups of four participants simultaneously performing the task on personalized but related reward environments (Fig. 1). This allows participants to make asocial generalizations based on private observations (using spatial correlations), and also integrate social information about the value of options they have not explored themselves (using social correlations).

Participants and design

We recruited *N=188* participants from Prolific in groups of four. They were assigned to their groups based on access time. After eliminating all groups with drop-out, the final sample size was *N=128* (mean age: 38.5 ± 12.7 SD; 44 females). The study was approved by the Ethics Committee of the University of Konstanz ('Collective learning and decision-making study'), and participants provided informed consent prior to participation. On average, participants spent 20.8 ± 0.5 minutes on the task and earned £ 7.19 ± 0.04.

Materials and procedure

The experiment was performed as a yoked online experiment in groups of four, where choices and reward outcomes were shown to the other participants in the group. After giving informed consent, participants were instructed that they were embarking on a scientific mission to collect salt samples from alien oceans on other planets. Each group member would be collecting a different salt, but the process generating salts was similar between all salts, explaining the social correlations. Their goal was to acquire as much salt as possible.

Participants were shown several fully revealed reward environments with the same specifications as in the experiments, and were required to pass a comprehension check to continue. Then, they entered a waiting room, which lasted up to 3 minutes from the first person entering. Participants were compensated for this additional waiting time (up to £0.2 per minute). If 3 participants hadn't joined a room after 3 minutes, the room was closed, and they were redirected to the post-experiment questionnaire directly. If a group of 4 had formed, the experiment started. Participants were presented with their own as well as the other participants' bandits with one tile revealed (Fig. 1a).

To generate the reward distributions for the socially correlated bandits, we first sampled parent distributions from a Gaussian Process (GP) prior to induce spatial correlations $f \sim \mathcal{GP}\left(0, k\left(\mathbf{x}, \mathbf{x}'\right)\right)$ with the variance fixed to $\sigma_\varepsilon^2 = 0.0001$. A radial basis function (RBF) kernel with length constant $\lambda = 2$ was used as the kernel function (see Eq. 2). Then, we generated candidate environments using the parent distribution as the mean function of the GP priors and the same RBF kernel for the covariance. After filtering candidates to be correlated with the parent environment by at least $r = 0.6$, we further selected environment sets that were correlated with each other at $r = 0.6 \pm 0.05$ (Fig. 1b).

Participants then had a search horizon of 14 trials within each round to gain as much reward as possible, which defined their bonus payment. After every choice, they would wait for all other group member's choices. Then, the outcomes of all other participants from the previous trial would be revealed simultaneously. To prevent one participant from holding up an entire group, a random choice would be generated if they took longer than 10 seconds. These cases (0.78% of choices) were excluded from analysis. This was repeated over 8 rounds. Whenever a round finished, participants were given feedback on what percentage of maximum possible payoff they achieved, and how this translated into their bonus payment. After the 8th round, they moved on to the post-experiment questionnaire.

Computational models

We first describe a Gaussian Process-Upper Confidence Bound (GP-UCB) agent as a baseline Asocial Learning model for the spatially correlated bandit task (Wu et al., 2018), which provides a rational solution in the absence of social information. We then introduce a number of candidate social learning models that build on the GP-UCB agent, each incorporating social information with different mechanisms. These social learning models integrate social information in increasingly complex ways, and Decision Biasing and Value Shaping were chosen based on previously existing literature, with Value Shaping having been found to best fit human behaviour (Najar et al., 2020).

Asocial Learning (AS). The baseline Asocial Learning model uses GP regression (Rasmussen & Williams, 2006) to make predictions about the expected reward and subjective uncertainty about all options on the grid. We then use an Upper Confidence Bound (UCB) sampling strategy and a softmax choice rule to convert the GP predictions into choice probabilities.

Conditioned on observations $\mathcal{D}_t = \{\mathbf{X}_t, \mathbf{y}_t\}$ of choices $\mathbf{X}_t$ (i.e. chosen arms of the bandit) and rewards $\mathbf{y}_t$, the GP provides posterior predictions about the reward for any target input $\mathbf{x}_*$ in the form of a Gaussian: $p(f(\mathbf{x}_*)|\mathcal{D}_t) \sim \mathcal{N}(m(\mathbf{x}_*), v(\mathbf{x}_*))$. Thus, the posterior predictions can be summarized in terms of their mean and variance, which are defined as:

$$\begin{aligned} m(\mathbf{x}_*) &= \mathbf{k}_{*,t}^{\top}(\mathbf{K}+\sigma_\varepsilon^2\mathbf{I})^{-1}\mathbf{y}_t \\ v(\mathbf{x}_*) &= k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_{*,t}^{\top}(\mathbf{K}+\sigma_\varepsilon^2\mathbf{I})^{-1}\mathbf{k}_{*,t}. \end{aligned} \tag{1}$$

$\mathbf{k}_{*,t} = k(\mathbf{X}_t, \mathbf{x}_*)$ is the covariance between observed and target inputs, and $\mathbf{K} = k(\mathbf{X}_t, \mathbf{X}_t)$ is the covariance between each pair of observed inputs. $\mathbf{I}$ is the identity matrix and σ_ε^2 is the observation variance, corresponding to assumed i.i.d. Gaussian noise on each reward observation. For the covariance function, we use a Radial Basis Function (RBF) kernel $k_{RBF}(\mathbf{x},\mathbf{x}')$ to describe the spatial correlation of rewards:

$$k_{RBF}(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\lambda^2}\right). \tag{2}$$

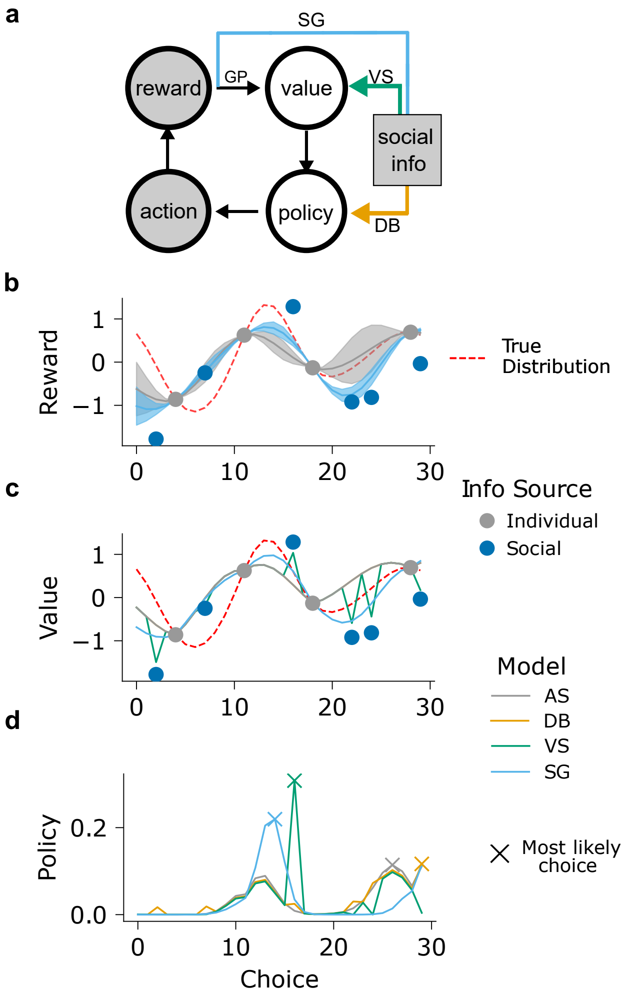

Figure 2: Models. **a**) Model overview. Individual reinforcement learning circuit (AS) and the stages at which different social models integrate social information (colored lines). Socially observable variables are grey, while unobservable variables are white. **b-d**) An illustrative 1D example of how models incorporate social information within the steps of the reinforcement learning circuit, where the x-axis is the discrete choice space. **b**) GP reward predictions (Eq. 1). Only SG integrates social information into the GP posterior, whereas the other models (only AS shown for ease of reading) generalize only private information. **c**) UCB values (Eq. 3). VS integrates social information into the value function proportional to its deviation from expected value. **d**) Softmax choice probabilities (Eq. 4). DB integrates social information into the policy based on choice frequency. Crosses mark the most likely choice for each model.

The length-scale parameter λ describes the decaying rate of covariance between two points as a function of distance, with higher λ capturing stronger spatial correlations. Thus, higher λ corresponds to broader generalization.

We then compute a *value function* V balancing exploiting high expected rewards with exploring uncertain options using UCB sampling:

$$V(\mathbf{x}) = m(\mathbf{x}) + \beta\sqrt{v(\mathbf{x})} \tag{3}$$

The uncertainty bonus β trades off the value of an option against the uncertainty of that estimate: high β values favor exploration, while low values favor exploitation.

Finally, we use a softmax to convert the value function into an individual learning *policy*:

$$\pi_{\text{ind}}(\mathbf{x}) \propto exp(V(\mathbf{x})/\tau) \tag{4}$$

The temperature parameter τ controls how deterministically the model follows the value function: the higher τ, the more random the choices become.

Decision Biasing (DB). We now present the simplest social learning model, incorporating social information into the policy as frequency-dependent copying (Toyokawa et al., 2019). We first define a pure social policy tracking observed choices in the previous trial, and selecting actions proportional to their frequency: $\pi_{\text{soc}}(x) \propto n_{x_{\text{soc}},t-1}$. Individual and social policies are then combined with the mixing parameter γ defining the relative contribution of the social policy:

$$\pi_{\text{DB}} = (1-\gamma)\pi_{ind} + \gamma\pi_{soc} \tag{5}$$

This results in an increased probability of choosing options previously chosen by other participants, regardless of outcome (Fig. 2d).

Value Shaping (VS). A relatively more complex strategy is to incorporate social information into the value function (Najar et al., 2020; Galef, 2013). Whereas previous studies assigned a generic bonus to socially observed choices since their outcomes were not shown to participants, we augment this approach to be value-sensitive using a simple prediction error update:

$$V(\mathbf{x}) = V_{x,ind} + \alpha(V_{x,soc} - V_{x,ind}) \tag{6}$$

Social observations that correspond to higher than expected payoffs ($V_{x,soc} > V_{x,ind}$) boost value estimates, while lower than expected payoffs ($V_{x,soc} < V_{x,ind}$) decrease them. The rate of this social boost/decrease is a function of the learning rate α. Figure 2c shows how the value function deviates from the asocial one in the direction of the social observation. We also considered a value-sensitive DB model, but elected to keep using the simpler, value-insensitive version, since value information did not improve performance.

Social Generalization (SG). We now provide a novel model which incorporates social information into the generalization process. Unlike the other models, it generalizes social information to surrounding options as well, but assumes it to be noisier, and thus less reliable, than individual information.

We model this by assigning a different value for the noise variance parameter based on whether an observation was individually or socially acquired:

$$\sigma^2_\varepsilon = \varepsilon_{\mathbf{ind}} + \delta_{soc} * \varepsilon_{\mathbf{soc}}, \tag{7}$$

ε_{ind} is the baseline level of noise, with indicator function $\delta_{\text{soc}} = 1$ for social observations, and 0 otherwise. Social noise parameter ε_{soc} is treated as a free parameter, where larger estimates correspond to less reliance on social information, relative to individual observations. This causes smaller updates to the mean and smaller reductions in uncertainty of the GP posterior for social observations compared to individual observations (Fig. 2b). When $\varepsilon_{\text{soc}} = 0$, social information is treated with the same weight as individual information.

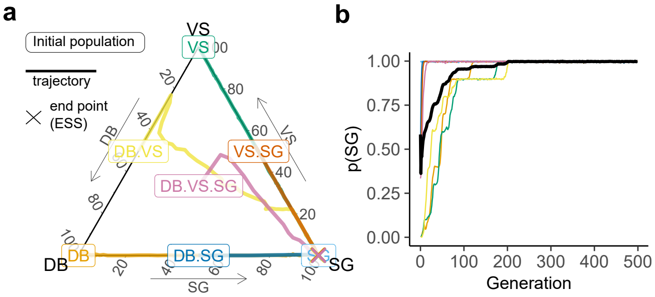

Figure 3: Evolutionary simulations. **a)** Initial populations (labels) and evolutionary trajectories (lines), terminating at the evolutionary stable strategy (ESS) after 500 generations (crosses). Initial populations always consisted of an equal ratio of included strategies (e.g. ”DB.VS” corresponds to 50% Decision Biasing and 50% Values Shaping). **b)** An overview of the proportion of SG agents in each generation, with the black line showing the overall mean across different initial populations.

Results

We first look at how the models perform in evolutionary simulations to find the normatively best model. The performance of social learners depends on the population they are in (Rogers, 1988), so it is necessary to evaluate the models' frequency-based fitness. Evolutionary simulations allow us to do this by covering all starting populations, instead of having to exhaustively evaluate every possible population composition. Then, we show behavioural results from the experiment and contrast with results from agent-based simulations. Finally, we present model fitting results to determine which model best fit human data.

Evolutionary Simulations

We used tournament selection in our evolutionary simulations (Tump, Wu, Bouhlel, & Goldstone, 2019): randomly sampled groups of four agents competed in one round of the socially correlated bandit task. The highest performing agent was selected to seed the next generation. Before the agents continued the process in the next generation, however, there was a set chance of mutations occurring. These could affect parameter values (p = .02 of adding Gaussian noise with $\sigma^2 = 0.2$), or change the agent's model completely (p = .002), which allowed models to invade populations that they were not originally a part of. GP-UCB parameters were sampled from prior distributions based on previous literature (λ and β from a lognormal distribution with mean≈ 0.54 and sd≈ 0.3, τ from a lognormal distribution with mean≈ 0.09 and sd ≈ 0.05; Wu et al., 2018). We selected plausible priors for the social parameters (α and γ uniformly between 0 and 1, and ε_{soc} from an exponential distribution with mean 2). Evolutionary trajectories starting from each initial populations of models were replicated 10 times.

Figure 3 shows the results of evolutionary simulations of

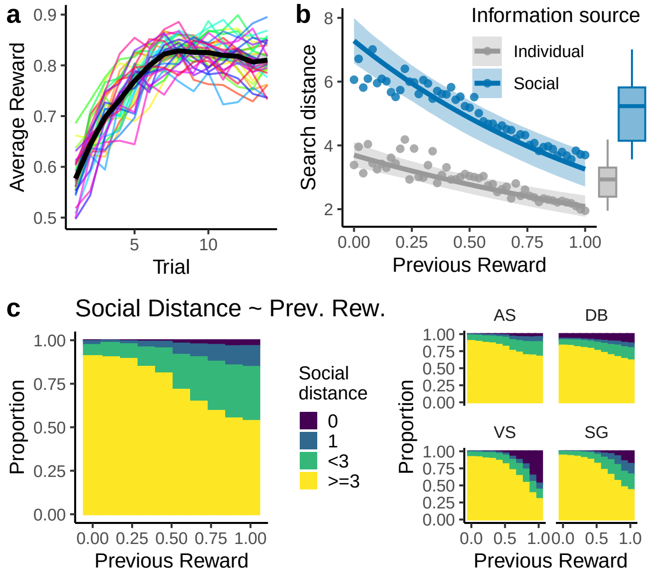

Figure 4: Behavioural results. **a**) Learning curves showing average reward over trial. Lines show group level averages, with the black line as the overall mean. **b**) Search distance over previous reward for individually and socially acquired information, analyzed hierarchically across participants (regression line and confidence interval). Each dot is population average over 50 equally spaced bins. **c**) Aggregate social distances over binned previous reward from participant data (left) and as predicted by the different models (right).

the competing social learning models. All initial populations — even ones that did not originally contain SG agents — evolved to be fully made up of SG agents. While the plot only shows social models, full simulations including AS had the same result, with the average $p(SG) = .998$ in the final generation. This clearly shows that SG is the normatively best model in our task. Additionally, the evolved $\varepsilon_{soc} = 3.29$ in the final generation suggests that using social information verbatim (i.e., $\varepsilon_{\text{soc}} = 0$) is not optimal.

Behavioural results

Next, we investigate signatures of human behaviour in the online experiment, focusing on aspects that would distinguish the different candidate models. As shown by the learning curves (Fig. 4a), participants succeeded at finding higher rewards throughout the trials.

To understand if and how humans use social and individual information differently, we next analyze search patterns as a function of previously observed reward. This is because each model make different predictions about how social information is used based on its value (no social info usage in AS; value-agnostic, frequency-based copying in DB; value-sensitive copying in VS; social generalization in SG). A rational agent will reduce their individual search distance as previous rewards increase due to the spatially correlated reward structure. When social information is used, one would expect a similar trend for social information.

Figure 4b shows a Bayesian linear regression of search distance (Euclidean distance between subsequent choices) as a function of the previous reward, split by whether the reward and distance are w.r.t. to social or individual observations. Social distance is calculated separately for each group member and choice.

Indeed, behaviour for individual information followed our prediction, with search distances decreasing as previous reward increases. The same trend could be seen for social observations: while social search distance is generally higher than individual, it also decreases over increasing reward. This is reflected in the significant negative regression weight of previous reward (-0.58; Highest Density Interval: [-0.74, -0.40]). The interaction effect between information source and previous reward is significant (-0.23; [-0.31, -0.16]), and implies participants may have been more sensitive to previous social than previous individual reward.

Breaking the social distances down further (Fig. 4c), we observe an increase of search in neighbouring (distance = 1) or near ($1 <$ distance < 3) options as previous reward increases, starting at low values but becoming more noticeable from around 0.5. There is hardly any exact imitation (distance = 0). Comparing this to the distributions expected based on our models (simulated data using the same parameter priors as the evolutionary simulations), it most closely resembles AS or SG, as DB and VS predict far higher levels of exact imitation. While the proportions of neighbouring and near searches most closely match SG, participants did not show enough exact imitation to fully replicate SG's overall distribution pattern. As a result, the participant data looks like a combination of AS (some increases in surrounding and near searches with hardly any exact imitation) and SG (numerical proportion of $1 \leq$ distance < 3 searches).

Model Results

Finally, we investigate model fits to human behavioural data. We used leave-one-out cross-validation to fit the models to participant data. Figure 5a shows the result of hierarchical Bayesian model selection (Rigoux, Stephan, Friston, & Daunizeau, 2014), where protected exceedance probability (pxp) describes the probability of a given model being predominant in the population (corrected for chance). We find that SG was the best model (pxp(SG) = .63) with $R^2 = 0.29 \pm$ 0.09 SD. AS also seems to be somewhat prevalent with pxp(AS) = .23, while neither DB nor VS were common.

This seeming mix of SG and AS in the population is reinforced by the parameter fits (Fig. 5b), with some ε_{soc}-values at the upper bound. Since higher assumed noise causes less deviation from the prior mean, higher ε_{soc}-values show lower (albeit not zero) reliance on social information. Participants' λ-estimates were significantly lower than the ground truth ($\lambda = 2$) with an average of $\hat{\lambda} = 1.11$ ($t(127) = -19.0$, $p < .001$, $d = 1.7$, $BF > 100$). Exploration bonus β had an average value of 0.29, and softmax temperature τ averaged 0.06. Participants with lower values of ε_{soc} (i.e., more weight for social information) achieved higher rewards on average ($r_\tau = -.28$, $p < .001$, $BF > 100$, Fig. 5c).

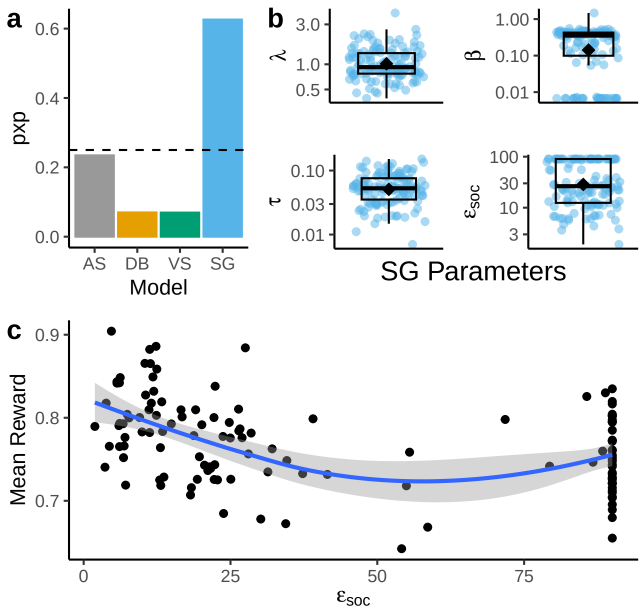

Figure 5: Model results. **a**) Bayesian model selection showing the probability of the best model across the population. Dashed line shows equal probability. **b**) Parameters estimates for SG. Diamonds show the mean, and each dot is a single participant. **c**) Average reward as a function of social noise parameter ε_{soc}, where lower noise corresponds to higher weight for social information. ε_{soc} estimates are bounded at 90, after which behavioural patterns become asymptotically asocial. Each dot is a participant, and the line and ribbon show the smoothed conditional mean.

Discussion

In this study, we sought to better understand how humans integrate social information when it is not perfectly applicable to their own situation. While this is a common scenario in our everyday experience, it is only sparsely researched. We presented a novel task, the socially correlated bandit, that operationalizes individual differences in social learning to achieve this goal.

Our evolutionary simulations show that Social Generalization (SG) is the best normative model in this setting. In a yoked experiment with groups of four participants, we found that participants were sensitive to social information in their search behavior. Our model comparison found that participants were predominantly best described by SG, although a non-negligible portion of the population used Asocial Learning (AS) instead. In contrast, we found far less evidence for simpler, heuristic models that directly imitated actions (Decision Biasing; DB) or used social information to modify value representations without generalization (Value Shaping; VS).

Limitations and future directions

The GP-UCB parameters we estimated differ from the ones found in previous literature (an average of $\hat{\lambda} = 0.5$, $\hat{\beta} = 0.51$) for the same ground truth environments in Wu et al., 2018). This may be a result of social learning, which was not present in previous studies: participants in our study appear to have undergeneralized individual information less, and showed less uncertainty-directed exploration. More accurate generalization may stem from increased information about the environment from other participants, while social learning may have partially replaced directed exploration. However, the median estimated noise parameter $\hat{\varepsilon}_{soc} = 26.4$ was still far higher than what the evolutionary simulations evolved towards $\varepsilon_{soc} = 3.29$, showing an underreliance on social information.

It has been repeatedly shown that humans underutilize social information in experimental settings, even when it is to their own detriment (Morin, Jacquet, Vaesen, & Acerbi, 2021). While not leading to optimal outcomes, it is still possible to perform the task without any social learning. While evolutionary simulations show that SG is the dominant strategy, it is more computationally complex, integrating 4x more data compared to AS. Formally, the degree of noise σ^2_ε can be related to Tikhonov regularization (Bishop, 1995), with higher ε_{soc} corresponding to a simpler, more regularized model. Thus, ignoring or underutilizing social information may be a resource-rational alternative (Bhui, Lai, & Gershman, 2021).

There is also the possibility that a subset of participants found the task setup (including the timer on the choices) too overwhelming to consider multiple sources of information. Some participants described only looking at other participants' grids at the very start of the experiment, and subsequently ignored them in favor of their own environment. While both VS and SG implicitly account for less reliance on social information as uncertainty is reduced, this may not have been enough to properly capture social information usage if it was limited to the very first trials. Future experiments could consider using shorter search horizons to further incentivize the use of social learning. Yet while this current design yields reliable model and parameter recovery (all $p(gen|fit) \geq .83$, $p(fit|gen) \geq .80$, and $r_\tau \geq .75$), shorter horizons may push the empirical limits of model estimation.

It is also important to note that this task is only designed to understand how social information, characterized by its positive correlation to individual information, would be integrated into decision-making. We make no claims about how information about the value others assign to an option could be inferred from their actions, but there is a plethora of work on the subject (Jara-Ettinger, 2019; Gweon, 2021).

Finally, since we studied groups of real participants, we had no control over the quality of social information available to participants. Potentially, there were groups with more or less valuable social cues, which would explain variance in how much social information was considered in individual decision-making. This could be circumvented in future work by having participants interact with artificial agents.

Conclusion

In summary, we find that the majority of humans used the normatively best model out of the ones we tested, implying that they are perfectly capable of taking social information with a grain of salt if the situation calls for it.

Acknowledgments

This work is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645, and funded by the DFG under Germany's Excellence Strategy – EXC 2117 – 422037984. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Alexandra Witt.

References

Analytis, P. P., Barkoczi, D., & Herzog, S. M. (2018). Social learning strategies for matters of taste. *Nature human behaviour*, *2*(6), 415–424.

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15–21.

Bishop, C. M. (1995). Training with noise is equivalent to tikhonov regularization. *Neural computation*, *7*(1), 108–116.

Charpentier, C. J., Iigaya, K., & O'Doherty, J. P. (2020). A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning. *Neuron*, *106*(4), 687–699.

FeldmanHall, O., & Nassar, M. R. (2021). The computational challenge of social learning. *Trends in Cognitive Sciences*, *25*(12), 1045–1057.

Galef, B. G. (2013). Imitation in animals: history, definition, and interpretation of data from the psychological laboratory. In *Social learning* (pp. 15–40). Psychology Press.

Gweon, H. (2021). Inferential social learning: how humans learn from others and help others learn. *Preprint at, https://doi. org/10*, *31234*.

Henrich, J. (2015). *The secret of our success*. Princeton University Press.

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

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

Laland, K. N. (2004). Social learning strategies. *Animal Learning & Behavior*, *32*(1), 4–14.

Laland, K. N. (2017). *Darwin's unfinished symphony*. Princeton University Press.

Morin, O., Jacquet, P. O., Vaesen, K., & Acerbi, A. (2021). Social information use and social information waste. *Philosophical Transactions of the Royal Society B*, *376*(1828), 20200052.

Müller-Trede, J., Choshen-Hillel, S., Barneron, M., & Yaniv, I. (2018). The wisdom of crowds in matters of taste. *Management Science*, *64*(4), 1779–1803.

Naito, A., Katahira, K., & Kameda, T. (2022). Insights about the common generative rule underlying an information foraging task can be facilitated via collective search. *Scientific Reports*, *12*(1), 1–12.

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

Rasmussen, C. E., & Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press: Cambridge, MA.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *Neuroimage*, *84*, 971–985.

Rogers, A. R. (1988). Does biology constrain culture? *American Anthropologist*, *90*(4), 819–831.

Toyokawa, W., Whalen, A., & Laland, K. N. (2019). Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, *3*(2), 183–193.

Tump, A. N., Wu, C. M., Bouhlel, I., & Goldstone, R. L. (2019). The evolutionary dynamics of cooperation in collective search. *bioRxiv*, 538447.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature human behaviour*, *2*(12), 915–924.

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

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2011). Receiving advice on matters of taste: Similarity, majority influence, and taste discrimination. *Organizational Behavior and Human Decision Processes*, *115*(1), 111–120.