Social Learning from Incomplete Information in a Dynamic Decision-Making Task

Alexandra F. Ortmann
Department of Psychology, Stony Brook University, NY 11794, USA

Christian C. Luhmann
Department of Psychology & Institute for Advanced Computational Science, Stony Brook University, NY 11794, USA

Abstract

The exploration-exploitation dilemma in dynamic decision-making scenarios is a notoriously hard problem to solve. Having a partner to potentially learn from might make it easier to balance exploration and exploitation. In the current study, we investigate the impact of social information (i.e., about others' exploration behavior vs. their rewards) and partner performance (optimal vs. random) on participants' behavior in a dynamic decision-making task that contains a learning trap. We find that observing the exploration behavior of an optimally choosing partner was detrimental to participants' overall performance and reduced participants' exploratory tendencies. In contrast, observing a random partner's exploration behavior stimulated participants' exploration, though this increase in exploration did not help participants to uncover the reward function. Following previous literature, a reinforcement learning model that contained eligibility traces was able to describe human behavior and helped to uncover potential mechanisms that could explain aspects of the findings.

Keywords: social learning; reinforcement learning; exploration-exploitation; dynamic decision-making

Introduction

The so-called exploration-exploitation dilemma (e.g., Cohen, McClure, & Yu, 2007; Sutton & Barto, 2018; for reviews see Hills et al., 2015; Mehlhorn et al., 2015), requires decision-makers to choose between the option they currently believe to be most valuable (exploitation) and other options they currently believe to be suboptimal but may turn out to be valuable (exploration). Past work investigating the exploration-exploitation dilemma has used a stylized task environment, commonly referred to as a multi-armed bandit, in which decision-makers attempt to maximize rewards by repeatedly selecting among multiple options. A broad range of such bandits can be designed and the complexity of the reward functions can vary dramatically. For example, reward functions can be dynamic, with the mean reward delivered by a given option changing over time (e.g., Daw, O'doherty, Dayan, Seymour, & Dolan, 2006; J. Li & Daw, 2011; Toyokawa, Saito, & Kameda, 2017; Toyokawa, Whalen, & Laland, 2019; for review see Gonzalez & Dutt, 2011) or contain learning traps, in which seemingly valuable options ultimately decrease earnings (Gureckis & Love, 2009b; Otto, Gureckis, Markman, & Love, 2009; Worthy, Gorlick, Pacheco, Schnyer, & Maddox, 2011). Humans often struggle to balance exploration and exploitation and to ultimately identify the optimal option, particularly in these complex, dynamic multi-armed bandits (Gonzalez, Lerch, & Lebiere, 2003; Gureckis & Love, 2009b, 2009a; Otto et al., 2009).

Exploration in Dynamic Decision-Making Tasks

To improve performance in these complex bandits, past work has investigated a plethora of interventions. For example, it has been argued that learning about the relationship of actions and rewards, which is particularly difficult in dynamic decision-making tasks, can be fostered by stimulating exploration (Gureckis & Love, 2009a; Tunney & Shanks, 2002). Studies have shown that exploration can be increased by experiencing losses instead of gains (Krueger, Wilson, & Cohen, 2017; Lejarraga, Hertwig, & Gonzalez, 2012; Lejarraga & Hertwig, 2017; Yechiam, Zahavi, & Arditi, 2015), which ultimately improves performance in a dynamic decision-making task (A. X. Li, Gureckis, & Hayes, 2021). Finally, another straightforward idea to foster learning is to provide different kinds of feedback, for example, about forgone rewards (A. X. Li et al., 2021; Otto & Love, 2010) or to adjust feedback presentation (Atkins, Wood, & Rutgers, 2002).

Social Learning in Dynamic Decision-Making Tasks

Learners often cannot only rely on their private information, but also observe others coping with the same learning problem. The act of learning from others—also referred to as social learning (e.g., Bandura, 1971; McElreath et al., 2005; Rendell et al., 2010, 2011; Whalen, Griffiths, & Buchsbaum, 2018)—has been shown to influence behavior in a variety of contexts (e.g., Boyd & Richerson, 1985, 2009; Harris, 2012; Laland, 2004; Molleman, Van den Berg, & Weissing, 2014; Rendell et al., 2010; Shafto, Goodman, & Griffiths, 2014). Bandura (Bandura, 1971; Bandura & Walters, 1977) highlighted that social learning offers potential advantages, but requires careful selection of social learning strategies. As a result, the act of social learning itself can be difficult (Kendal et al., 2018; Laland, 2004). This difficulty is composed of at least two separate challenges.

A first challenge is who to learn from. Intuitively, the ideal scenario involves learning from partners that act optimally. Gonzalez (2005) placed participants in a complex learning environment and found that observing the behavior of an expert improved participants' performance. In contrast, conforming to others that are not necessarily behaving optimally seems potentially fraught (McElreath et al., 2005; Toyokawa

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

et al., 2019). Therefore, whether social learning is beneficial strongly depends on those being learned from.

A second challenge is that social learners may have limited information about others. Some information such as other's goals and intentions might be inherently unobservable, and even observable information (e.g., rewards and choices) may not be available to social learners. Prior work suggests that access to others choices is generally beneficial (Vélez & Gweon, 2019; Whalen et al., 2018), though the benefit may vary with task complexity. Access to the rewards others are earning may increase exploration, which can be helpful in difficult tasks but may be detrimental in simpler tasks (Nedic, 2011). Counter-intuitively, more social information is not always better. Toyokawa, Kim, and Kameda (2014), for example, found that access to others' evaluations, in addition to choices, can actually be detrimental to individuals' performance.

The Present Study

As we have reviewed, people struggle to identify and choose the optimal option in dynamic decision-making tasks. Social learning is often successful when others' behavior is optimal (or at least superior to one's own). However, in everyday life people often have only limited access to information from other social sources, and the benefit of social learning depends on what types of information people have access to.

In the current study, we asked whether participants contending with a complex decision-making task can use limited information about the behavior of a partner to improve their own performance. Specifically, we were interested in how observing others' explorative tendencies (i.e., how often they are switching, or alternating among options) without seeing the choices themselves influences participants' choices and, ultimately, their overall performance. Observing others' exploratory choice behavior could be beneficial as exploration has been suggested to be a successful strategy in complex environments where optimal strategies may be difficult to detect. However, it could be detrimental because participants are only able to observe a specific facet of their partner's behavior and it is, presumably, not the information participants most desire. As a comparison, other participants were given access to the rewards their partners were earning, again without seeing the choices that led to those rewards. Further, as discussed earlier, the value of social information depends on others' performance. Therefore, partners were algorithmic (rather than other participants) and either performed optimally or switched randomly. Overall, there are two variables *Information Type* (Switch vs. Reward) and *Partner Type* (Random vs. Optimal) resulting in four conditions that are called switch-random, switch-optimal, reward-random and reward-optimal.

Intuitively, participants observing an Optimal partner should perform better than participants observing a Random partner. Further, previous findings (Nedic, 2011) suggest that information about rewards may be more helpful than information about partners' choice behavior. However, the advantages social information might provide should critically depend on the interaction between Partner and Information Type sometimes leading to surprising patterns. Therefore, we are specifically interested in two hypotheses:

1. Observing actions from an Optimal partner is not always helpful: Participants in the switch-optimal condition will alternate less and choose the optimal option less than participants in the switch-random and reward-optimal conditions.

2. Observing actions from a Random partner is helpful: Participants in the switch-random condition will alternate most and choose the optimal option more often than the reward-random condition.

Method

This study was approved by the Institutional Review Board at Stony Brook University.

Participants

Seventy-two undergraduate students (18 per condition) participated in exchange for partial course credit (age M = 19.99, 64% female). One additional participant was excluded due to a technical error that caused the task to terminate prematurely.

Task

The task builds on the "Farming on Mars" task (Gureckis & Love, 2009a, 2009b), which has been modified for different purposes (Cooper, Worthy, Gorlick, & Maddox, 2013; Cooper, Worthy, & Maddox, 2016; Otto et al., 2009; Otto & Love, 2010; Worthy et al., 2011; Worthy, Otto, & Maddox, 2012) and is a member of a larger family of dynamic decision-making tasks (Gonzalez, Fakhari, & Busemeyer, 2017; Herrnstein, 1991; Herrnstein & Prelec, 1991; Rahmandad, Denrell, & Prelec, 2021; Sims, Neth, Jacobs, & Gray, 2013; Tunney & Shanks, 2002). Participants in the Farming on Mars task are told that two robots have been sent to Mars in order to produce oxygen. Participants are asked to repeatedly choose which of the two robots should be used to produce oxygen. Once selected, the amount of oxygen produced by the robot is reported to the participant, allowing participants to learn about the capability of the selected robot. Forgone rewards (oxygen units) are not presented. The participant's goal is to maximize the total amount of oxygen produced.

The task appears to be a standard multi-armed bandit task often used in reinforcement learning settings (Cohen et al., 2007; Daw et al., 2006; Sutton & Barto, 2018). The robots represent the "arms" or options and the oxygen represents the reward to be maximized. Unlike conventional multi-armed bandit tasks, however, the reward function is more complex, largely due to the dynamic nature of the rewards. Unbeknownst to participants, the two robots not only offered systematically different rewards on each trial, but also had qualitatively different impact on future rewards.

$$R_i = x_i + 1000 * (h/10) + N(\mu, \sigma) \quad (1)$$

We distinguish between the short-term option, $i = 1$, and the long-term option, $i = 2$. The reward, R, offered by each option is, in part, determined by a fixed component x that varied between the two options. For the short-term option, this fixed component was relatively larger, x_1=900, and for the long-term option, this fixed component was relatively smaller, x_2=400. Thus, on any given trial, the reward offered by the short-term option was larger (by 500 units) than the reward offered by the long-term option. However, the total reward also depended on a dynamic component, h, which captured the number of times the long-term option was selected on the previous 10 trials. The existence of h introduces long-range dependencies into the reward structure and is what makes the task dynamic. Finally, a small amount of normally-distributed noise (μ=0, σ=50) was added to the reward on each trial. In this way, the short-term option is more valuable in the short-term, but selecting the long-term option maximizes long-term earnings. Therefore, cumulative earnings are maximized by consistent selection of the long-term option.

Participants were also presented with information about the behavior of a partner. Participants were told that the partner was in the same situation as they were and that details of the partner's behavior would be made available incrementally throughout the task. In actuality, the behavior of the partner reflected one of two simple algorithms. The behavior of the Optimal partner followed a simple rule of choosing the long-term option with probability 0.95 on each trial. Therefore, the Optimal partners rarely switched between options and tended to receive large rewards. The behavior of the Random partner followed a simple rule in which choices tended to alternate from trial to trial. Specifically, the Random partner selected the option selected on the previous trial with probability 0.25 (i.e., switching with probability 0.75). Therefore, the Random partner switched between options quite frequently and tended to receive moderate rewards.

Participants were only given partial information about their partner's behavior. Specifically, participants in the Switch condition were presented with information about whether their partner alternated (switched) or repeated their previous selection. Participants in the Reward condition were presented with information about the rewards their partner earned. In neither condition did participants have access to information about the options selected by their partner (i.e., long-term or short-term).

Procedure

Participants completed the experiment in person. After providing consent, participants made a sequence of 175 choices. Each trial consisted of a choice phase, during which the options were presented and subjects could make their selection via a keyboard key press. After a choice was made, the selected choice was highlighted for 1 second. The outcome of the choice (i.e., number of oxygen units generated) was then presented for 1.5 seconds. Information regarding the most recent choices (of both the participant herself and the partner) was presented for an additional 1.5 seconds. The next trial began immediately thereafter.

Historical information regarding the previous five trials remained on screen at all times. This information was comprised of the last five rewards earned by the participant as well as the last five trials-worth of information about the partner (alternations or rewards depending on condition). Which side of the screen the long-term and short-term options were presented on and the color of the robots was counter-balanced across participants.

Results

The goal of the current study was to better understand the effects social learning might have on a notoriously difficult dynamic decision-making task. Participants observed either the exploration tendencies (i.e., alternations) or the rewards earned by either a Random or Optimal partner.

In the following, we focus on two different outcomes of interest. First, the alternation rate of participants is of interest as exploration has been suggested as a critical strategy to overcome learning traps and might directly relate to the alternation rate of their partner. Second, the proportion of long-term choices is analyzed as in previous studies (Gureckis & Love, 2009a, 2009b).

Descriptive Results

Over 175 trials participants chose the long-term option on M = 51.21 (29.26%; SD = 22.3%) trials, generating $M = 183,756$ (SD = 18,188.22) total oxygen units, and switching between options (i.e., alternating) on M = 35.54 trials (M = 20.31%, SD = 12.86%). Figures 1a and 1b illustrate the alternation rate and proportion of long-term choices for each condition. As is typical for multi-armed bandit tasks (e.g., Gonzalez & Dutt, 2011; Lejarraga & Hertwig, 2017), participants explored more at the beginning than at the end of the trial sequence (see Figure 1a). Further, it seems that participants who received information about the alternation behavior of an Optimal partner alternated least and participants who received information about the alternation behavior of a Random partner alternated most. Compared to other groups, participants observing their Optimal partners' rewards ended up choosing the long-term option most.

Bayesian Model: Group Differences

Model Specification Each outcome is assumed to be described by a linear function with a subject-level intercept, three nominal factors, and noise (error). The intercept was modeled as a random effect. Each participant-specific intercept was assumed to be drawn from an overarching normal distribution (μ=0, σ=.5), which acted as a weakly informative hyper parameter. The nominal factors were effect coded, with the first factor representing the deflection from a baseline due to *Information Type* (-1=reward, 1=alternation), the

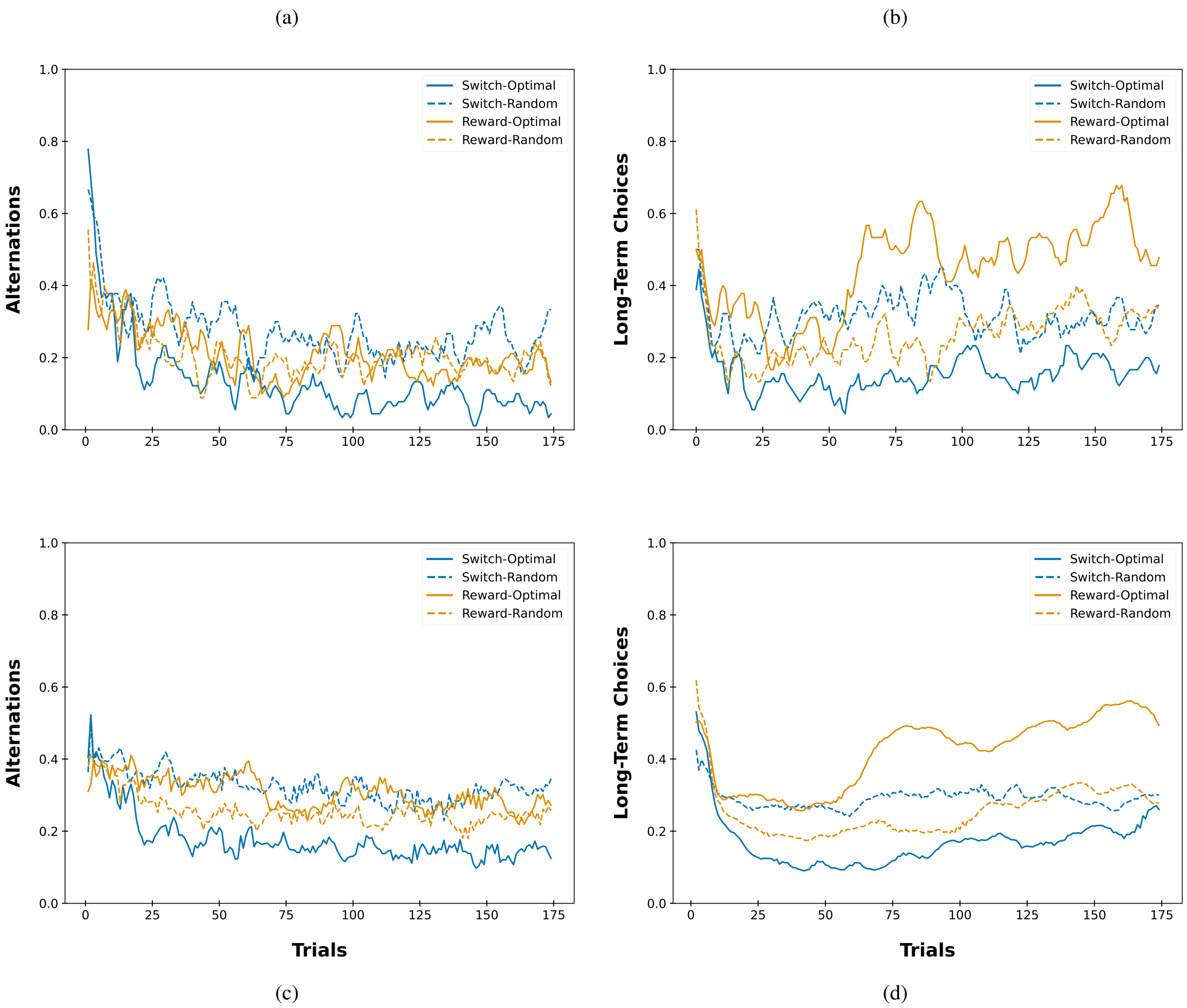

Figure 1: (a) & (b) Behavioral data plotted by experimental condition; time courses smoothed using rolling window of five trials; Proportion of (a) alternations, (b) long-term choices. (c) & (d) Results of model simulation using best fitting parameter values and participants' actual choice histories. Probability (c) to alternate, (d) of long-term choice.

second factor representing the deflection due to *Partner Behavior* (-1=random, 1=optimal), and the last factor represents the interaction of the two. Coefficients associated with these three factors were each informed by separate, weakly informative priors that were assumed to be normally distributed (μ=0, σ=1) and were modeled as fixed effects. The predicted outcome was assumed to be normally distributed and was given a prior Half-Cauchy distribution to reflect noise in the data. Standardized regression coefficients of the hierarchical Bayesian mixed-effects regression model are reported. The MCMC process generated stable estimates of the posterior distribution and the traceplots explored the parameter space reasonably well.

Results First, we analyzed whether participants' alternation behavior was related to whether they saw a partner's alternation behavior or their rewards (see regression summary statistics in Table 1). Overall, Information Type did not strongly influence alternation ($p(\hat{\beta}_1 > 0)$=.412, $M_{\hat{\beta}_1}$=-.02). In contrast, Partner Type did ($p(\hat{\beta}_2 > 0)$=.009, $M_{\hat{\beta}_2}$=-.26). When participants collaborated with an Optimal partner, they alternated less than if they collaborated with a Random partner. Further, it is very likely that there is an interaction between Information Type and Partner Type ($p(\hat{\beta}_{1x2} > 0)$ = .005). To better interpret these results, we investigated unstandardized simple effects.

In line with the descriptive data presented in Figure 1a,

Table 1: Regression Summary Statistics

	Mean	SD	p(>0)	HDI 3%	HDI 97%
Alternations					
Information	-.024	.111	.412	-.231	.189
Partner	-.263	.110	.009	-.475	-.061
Info. x Partner	-.287	.110	.005	-.498	-.081
Long-Term Choices					
Information	-.261	.106	.008	-.461	-.065
Partner	.029	.106	.608	-.171	.225
Info. x Partner	-.383	.107	.000	-.583	-.180

Note. p(>0) represents posterior probability that effect is >0 (values can range from 0-1, 0 indicating potential negative effect, 1 indicating potential positive effect). *HDI* columns represent highest density interval for the corresponding coefficient.

participants in the Reward condition alternated to a similar extent, independent of whether they collaborated with a Random ($\hat{M}_{alternation}$ = .20) or Optimal ($\hat{M}_{alternation}$ = .21) partner (p([reward-optimal > reward-random])= .560). In contrast, participants in the Switch condition alternated more when they collaborated with a Random partner ($\hat{M}_{alternation}$=.27) than if they collaborated with an Optimal partner ($\hat{M}_{alternation}$=.13; p([switch-optimal > switch-random])< .001).

Next, we investigated participant's choice behavior (see Table 1). Information Type did influence how often participants chose the long-term option ($p(\hat{\beta}_1 > 0)$=.008, $M_{\hat{\beta}_1}$=-.26). Participants likely developed a stronger preference for the long-term option when they observed a partner's rewards rather than a partner's alternation behavior. In contrast, a partner's performance (i.e., Partner Type) did not have any overall impact on the proportion of participants' long-term choices ($p(\hat{\beta}_2 > 0)$=.608, $M_{\hat{\beta}_2}$=.029). However, similar to participants' alternation behavior, there is a strong interaction ($M_{\hat{\beta}_{1x2}}$=-.38, $p(\hat{\beta}_{1x2}>$ 0)<.001).

With near certainty participants who worked with an Optimal partner preferred the long-term option more when seeing their partner's rewards ($\hat{M}_{Long-term\ choice}$=.44) than when seeing their partner's alternation behavior ($\hat{M}_{Long-term\ choice}$=.16; p([reward-optimal > reward-random])= .999). In contrast, participants chose the long-term option to the same extent regardless of whether they saw a Random partner's rewards ($\hat{M}_{Long-term\ choice}$=.26) or alternations ($\hat{M}_{Long-term\ choice}$=.31, p([switch-optimal > switch-random])= .207).

Reinforcement Learning Model: Describing and Predicting Behavior

When employing repeated decision-making tasks, RL models are often used to describe behavior. Gureckis and Love (2009b) tested a range of RL models in the Farming on Mars task. In the following, we will use one of these, consisting of a Q-learning update rule, a softmax choice function, and an eligibility trace (ET). This model is of particular interest because the ET permits memory for past choices which is particularly critical in the current task.

We took a two step approach to determine whether the model could recreate human behavior. First, the model was fitted to the choice data collected in the experiment and the parameter estimates analyzed. Second, we simulated alternation and choice behavior on individual trials by using the best-fitting parameter values and the participant's sequence of choices over prior trials. This approach reproduces the parameter estimation routine. Together, these evaluations allow us to investigate how and how well the ET model accounts for human behavior in this task and provide insights into mechanisms that can and cannot explain aspects of the findings.

Model Definition After every trial, t, the model updates the subjectively expected outcome, Q_j, associated with each option, j, based on the difference between the subjectively expected and experienced outcome, r, also called prediction error. These adjustments are modulated by a learning rate, $0 \leq \rho \leq 1$. The ET model includes a trace, λ, that captures how often each option has been chosen in the past and uses this trace to update beliefs of recently chosen options (Bogacz, McClure, Li, Cohen, & Montague, 2007; Gureckis & Love, 2009b). How rapidly memory for past actions fades is controlled by a decay parameter, τ. Greater values of τ represent a faster decay (i.e., outcomes are attributed to choices in the recent rather than far past). A decay parameter of τ=1 reduces the ET model to the conventional RL model.

$$Q_{j,\,t+1} = Q_{j,\,t} + \rho\lambda_{j,t}[r_{j,t} - Q_{j,t}] \tag{2}$$

$$\lambda_{j,\,t} = \begin{cases} (1-\tau)\lambda_{j,t-1} + 1 & \text{if } j \text{ is selected} \\ (1-\tau)\lambda_{j,t-1} & \text{otherwise} \end{cases} \tag{3}$$

The model feeds these subjectively expected outcomes into a softmax choice function (Luce, 1959; Sutton & Barto, 2018) which generates probabilistic preferences and, ultimately, a dichotomous choice on each trial. θ is a free parameter that indicates the stochasticity of participants' choices.

$$p_{j,t} = \frac{e^{\theta Q_{j,t}}}{e^{\theta Q_{1,t}} + e^{\theta Q_{2,t}}} \tag{4}$$

Estimation Applying a global optimization algorithm (Virtanen et al., 2020; Wales & Doye, 1997) to the behavioral data described above, we generated participant-level estimates of ρ, τ, and θ.

Estimated parameter values suggested that the manipulations of Partner and Information Type influenced distinct aspects of the learning process (see Table 2). Partner Type influenced the learning rate for those receiving switch information, with participants in the switch-random condition exhibiting the most rapid updating (M_ρ = .527), likely supporting the increased alternation observed in that condition. In

Table 2: Mean Parameter Estimates.

	Learning Rate (ρ)	ET decay (τ)	Determinism ($100 * \theta$)
Reward-Optimal	.410	.596	.303
Reward-Random	.424	.760	.414
Switch-Optimal	.208	.632	.563
Switch-Random	.527	.628	.384

contrast, Partner Type influenced the decay rate for those receiving reward information. Estimates of the ET decay rate were greatest in the reward-random condition (M_τ = .760) and smallest in the reward-optimal condition (M_τ = .596). It is also notable that estimated decay values implied that *all* participants attributed outcomes to choices made on earlier trials—a central mechanism of the task environment—to some extent.

Qualitative Check When participants' specific choice histories were used, the ET model reproduced the general trends found in human alternation and choice behavior (see Figures 1c and 1d). In the reward-optimal condition, participants were expected to choose the long-term option more often, especially after trial ~50. Further, participants in the switch-optimal condition were expected to choose the long-term option the least often and also alternate least. In the two Random partner conditions, the models chose the long-term option to roughly the same extent, especially towards the end of the trial sequence.

Overall, the ET model seems capable of reproducing most of the group differences found in the experiment. Especially the model's ability to reproduce alternations despite using parameters fitted to choice behavior, suggests that the model might be a good description of the underlying mechanism.

Discussion

Overall, we found that both Partner Type (Random vs. Optimal) and social Information Type (Switch vs. Reward) impacted participants' alternation behavior and choices. First, participants with access to an Optimal partner's rewards were the only participants that consistently developed any preference for the long-term option and, consequently, earned more than participants in any other condition. This finding is in line with previous results (Nedic, 2011), which have demonstrated that access to others' rewards can be more helpful than access to others' choice behavior. However, in contrast to Gonzalez (2005) and potentially intuition, learning from an Optimal partner was not always advantageous. If participants only had access to an Optimal partner's alternation behavior, they explored less than, and performed worse than, any other participants. This seemingly counterintuitive and divergent finding can be attributed to only having access to incomplete information. Finally, participants' alternation behavior revealed that participants who had access to a Random partner's alternation behavior alternated more than any other participants and participants who had access to an Optimal partner's alternation behavior, alternated less than any other participants. These results indicate that participants imitated their partner's behavior, at least to some extent. However, the increase in exploration evoked by observing a Random partner's alternation behavior was not sufficient to lead participants out of the learning trap and toward more optimal performance.

We analyzed our behavioral data using an RL model that included eligibility traces. The model yielded sensible parameter estimates, and was able to account for the behavioral results for both long-term choice preferences and alternations. Models exhibited less alternation and updated beliefs more slowly (i.e., lower learning rate) in the switch-optimal condition than in any other group. We then used participants' specific choice sequence to generate single choices. The resulting simulated data revealed that the model was able to recreate the preferences found in behavioral data. Overall, the parameter estimates are moderately consistent with the data, and were sufficient to reproduce the relevant aspects of participants' behavior. Further, it seemed like the experimental manipulations, especially with respect to the type of partner information, might impact different steps within the learning process. For example, participants observing partner's alternation behavior showed different learning rates dependent on what type of partner they collaborated with. Participants who saw partner's rewards showed differences in the ET decay, but not the learning rate. Given these findings, the models' ability to serve as a description of participant's learning in an as-if capacity should be recognized.

Despite the performance of the ET model, the primary opportunities for future work lie in model development. Although existing social learning models (e.g., McElreath et al., 2005; Najar, Bonnet, Bahrami, & Palminteri, 2020; Nedic, 2011; Toyokawa et al., 2017) can be adapted, they cannot be used as-is as they rely on integrating information about partners' choice behavior, which our participants did not have access to. Another interesting question is whether social learning in this task is advantageous relative to individual learning. In addition to a 'no social learning' control condition, a condition in which participants have access to a partner's alternation behavior as well as their rewards might be of further interest. Such a design would permit inferences about which combination of information types offers the best opportunities to learn from social sources.

Deciding between exploring and exploiting different options is not an easy task by itself. Having access to information about how others make this trade-off might seem to help weigh exploration and exploitation. However, the current study shows that this notion crucially depends on the type of information and performance of the partner in non-trivial ways. Senselessly emulating a partner's behavior can have beneficial and detrimental effects, but does not seem to be the right answer either way.

References

Atkins, P. W., Wood, R. E., & Rutgers, P. J. (2002). The effects of feedback format on dynamic decision making. *Organizational Behavior and Human Decision Processes*, *88*(2), 587–604.

Bandura, A. (1971). Vicarious and self-reinforcement processes. *The Nature of Reinforcement*, *228278*.

Bandura, A., & Walters, R. H. (1977). *Social learning theory* (Vol. 1). Englewood Cliffs Prentice Hall.

Bogacz, R., McClure, S. M., Li, J., Cohen, J. D., & Montague, P. R. (2007). Short-term memory traces for action bias in human reinforcement learning. *Brain Research*, *1153*, 111–121.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary trocess*. The University of Chicago Press.

Boyd, R., & Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1533), 3281–3288.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942.

Cooper, J. A., Worthy, D. A., Gorlick, M. A., & Maddox, W. T. (2013). Scaffolding across the lifespan in history-dependent decision-making. *Psychology and Aging*, *28*(2), 505.

Cooper, J. A., Worthy, D. A., & Maddox, W. T. (2016). Information about foregone rewards impedes dynamic decision-making in older adults. *Aging, Neuropsychology, and Cognition*, *23*(1), 103–116.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Gonzalez, C. (2005). Decision support for real-time, dynamic decision-making tasks. *Organizational Behavior and Human Decision Processes*, *96*(2), 142–154.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological Review*, *118*(4), 523.

Gonzalez, C., Fakhari, P., & Busemeyer, J. (2017). Dynamic decision making: Learning processes and new research directions. *Human Factors*, *59*(5), 713–721.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Gureckis, T. M., & Love, B. C. (2009a). Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology*, *53*(3), 180–193.

Gureckis, T. M., & Love, B. C. (2009b). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition*, *113*(3), 293–313.

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Harvard University Press.

Herrnstein, R. J. (1991). Experiments on stable suboptimality in individual behavior. *The American Economic Review*, *81*(2), 360–364.

Herrnstein, R. J., & Prelec, D. (1991). Melioration: A theory of distributed choice. *Journal of Economic Perspectives*, *5*(3), 137–156.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., Group, C. S. R., et al. (2015). Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences*, *19*(1), 46–54.

Kendal, R. L., Boogert, N. J., Rendell, L., Laland, K. N., Webster, M., & Jones, P. L. (2018). Social learning strategies: Bridge-building between fields. *Trends in Cognitive Sciences*, *22*(7), 651–665.

Krueger, P. M., Wilson, R. C., & Cohen, J. D. (2017). Strategies for exploration in the domain of losses. *Judgment and Decision Making*.

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

Lejarraga, T., & Hertwig, R. (2017). How the threat of losses makes people explore more than the promise of gains. *Psychonomic Bulletin & Review*, *24*(3), 708–720.

Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, *124*(3), 334–342.

Li, A. X., Gureckis, T. M., & Hayes, B. (2021). Can losses help attenuate learning traps? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Li, J., & Daw, N. D. (2011). Signals in human striatum are appropriate for policy update rather than value prediction. *Journal of Neuroscience*, *31*(14), 5504–5511.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, *66*(2), 81.

McElreath, R., Lubell, M., Richerson, P. J., Waring, T. M., Baum, W., Edsten, E., … Paciotti, B. (2005). Applying evolutionary models to the laboratory study of social learning. *Evolution and Human Behavior*, *26*(6), 483–508.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., … Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191.

Molleman, L., Van den Berg, P., & Weissing, F. J. (2014). Consistent individual differences in human social learning strategies. *Nature Communications*, *5*(1), 1–9.

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.

Nedic, A. (2011). *Models for individual decision-making with social feedback* (Unpublished doctoral dissertation). Princeton University.

Otto, A. R., Gureckis, T. M., Markman, A. B., & Love, B. C. (2009). Navigating through abstract decision spaces: Evaluating the role of state generalization in a dynamic decision-making task. *Psychonomic Bulletin & Review*, *16*(5), 957–963.

Otto, A. R., & Love, B. C. (2010). You don't want to know what you're missing: When information about forgone rewards impedes dynamic decision making. *Judgment and Decision Making*, *5*(1), 1.

Rahmandad, H., Denrell, J., & Prelec, D. (2021). What makes dynamic strategic problems difficult? evidence from an experimental study. *Strategic Management Journal*, *42*(5), 865–897.

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

Rendell, L., Fogarty, L., Hoppitt, W. J., Morgan, T. J., Webster, M. M., & Laland, K. N. (2011). Cognitive culture: theoretical and empirical insights into social learning strategies. *Trends in Cognitive Sciences*, *15*(2), 68–76.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Sims, C. R., Neth, H., Jacobs, R. A., & Gray, W. D. (2013). Melioration as rational choice: sequential decision making in uncertain environments. *Psychological Review*, *120*(1), 139.

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

Toyokawa, W., Kim, H.-r., & Kameda, T. (2014). Human collective intelligence under dual exploration-exploitation dilemmas. *PloS one*, *9*(4), e95789.

Toyokawa, W., Saito, Y., & Kameda, T. (2017). 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.

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.

Tunney, R. J., & Shanks, D. R. (2002). A re-examination of melioration and rational choice. *Journal of Behavioral Decision Making*, *15*(4), 291–311.

Vélez, N., & Gweon, H. (2019). Integrating incomplete information with imperfect advice. *Topics in cognitive science*, *11*(2), 299–315.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Wales, D. J., & Doye, J. P. (1997). Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, *101*(28), 5111–5116.

Whalen, A., Griffiths, T. L., & Buchsbaum, D. (2018). Sensitivity to shared information in social learning. *Cognitive Science*, *42*(1), 168–187.

Worthy, D. A., Gorlick, M. A., Pacheco, J. L., Schnyer, D. M., & Maddox, W. T. (2011). With age comes wisdom: Decision making in younger and older adults. *Psychological Science*, *22*(11), 1375–1380.

Worthy, D. A., Otto, A. R., & Maddox, W. T. (2012). Working-memory load and temporal myopia in dynamic decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1640.

Yechiam, E., Zahavi, G., & Arditi, E. (2015). Loss restlessness and gain calmness: durable effects of losses and gains on choice switching. *Psychonomic Bulletin & Review*, *22*(4), 1096–1103.