Modeling Social Learning Through Demonstration in Multi-Armed Bandits

Julio Martinez[1,*], **Michael C. Frank**[1], **Nick Haber**[2,3]
[1]Department of Psychology, Stanford University
[2]Graduate School of Education, Stanford University
[3]Department of Computer Science, Stanford University
[*]juliomz@stanford.edu

Abstract

Humans are efficient social learners who leverage social information to rapidly adapt to new environments, but the computations by which we combine social information with prior knowledge are poorly understood. We study social learning within the context of multi-armed bandits using a novel "asteroid mining" video game where participants learn through active play and passive observation of expert and novice players. We simulate human exploration and social learning using naïve versions of Thompson and Upper Confidence Bound (UCB) solvers and hybrid models that use Thompson and UCB solvers for direct learning together with a multi-layer perceptron to estimate what should be learned from other players. Two variants of the hybrid models provide good, parameter-free fits to human performance across a range of learning conditions. Our work shows a route for integrating social learning into reinforcement learning models and suggests that human social learning conforms to the predictions of such models.

Keywords: social learning, multi-armed bandits, reinforcement learning

Introduction

Social learning is fundamental to human cognitive development (Harris, 2012). It enables individuals to assimilate knowledge efficiently through exposure to social information (Rosenberg & Vieille, 2019) especially when firsthand experience is costly or when environments are complex. Additionally, social learning plays a key role in development, skill acquisition, emotional development, the acquisition of language, and the transmission of cultural knowledge (Duranti, Ochs, & Schieffelin, 2014; Giraldeau, Caraco, & Valone, 1994).

How is socially-learned information integrated with information learned from direct experience? The experimental and cognitive neuroscience literature suggests that social learning is similar to reward-based self-exploratory learning, with mechanisms for social evaluation that rely on associative processes to make predictions about social partners, and that these social predictions are integrated with self-exploration prior to decision making (Behrens, Hunt, Woolrich, & Rushworth, 2008; Behrens, Hunt, & Rushworth, 2009; Olsson, Knapska, & Lindström, 2020; Charpentier & O'Doherty, 2018). Moreover, learners evaluate the expertise of their social partners – even young children can learn selectively from more accurate and more expert models (Boorman, O'Doherty, Adolphs, & Rangel, 2013; Harris, 2012). On the other hand, some evidence demonstrates that humans are disproportionately influenced by social information (Muthukrishna, Morgan, & Henrich, 2016; Laland, 2004) and that they adjust quantity estimates after receiving information from social partners (Molleman, Kurvers, & van den Bos, 2019) despite uncertainty about their partner's estimates.

Although there is a rich body of experimental work demonstrating that humans are indeed social learners and that social information enables complex behaviors, formal computational theories for explaining how humans leverage social information are less well-developed (FeldmanHall & Nassar, 2021). For example, how does the computational process of social learning change when your teacher is an expert compared with a novice, or when the social partner is operating under uncertainty? Some theories of social cognition suggest that humans form generative models that allow them to simulate the mental states of social partners such as their current world understanding, goals and intentions in order to better interact with and learn from them (Baker, Saxe, & Tenenbaum, 2009; Vélez & Gweon, 2021; Charpentier & O'Doherty, 2018).

Inspired by this body of work we choose to study the problem of social information integration under uncertainty in the context of *multi-armed bandits*. Multi-armed bandits are simple choice problems where a decision maker repeatedly selects one of several choices by selecting a bandit "arm". In particular, Bernoulli multi-armed bandits are bandit problems where each bandit arm i has a payoff probability p_i of 1 or 0. Bernoulli bandits have been studied extensively in humans to understand how humans search for rewards under uncertainty (Burtini, Loeppky, & Lawrence, 2015; Schulz, Franklin, & Gershman, 2020; Schulz, Konstantinidis, & Speekenbrink, 2015; Stojic, Analytis, & Speekenbrink, 2015; Gray, Zhu, Arigo, Forman, & Ontañón, 2020). In addition, there has been an extensive body of work on the optimal strategies for solving multi-armed bandits (Kuleshov & Precup, 2014). Therefore, they are a good system for investigating the process of integrating direct exploration and social learning in detail.

Prior work has used small-scale bandit problems to study social learning. In one study using a three-armed bandit task, learners processed observed outcomes from another player in a similar manner to their own observations (Adrian, Siddharth, Baquar, Jung, & Deák, 2019). Another study, in-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

volving a two-armed bandit task with decision cues and advice from a confederate, found that incorporating both the advice of the confederate and the true probabilities of the bandit resulted in more accurate predictions of human decisions (Behrens et al., 2008). These studies suggest a complex interplay between direct experience and social information in these learning problems.

Our goal in the current work is to use multi-armed bandits to study the integration of self-exploration knowledge with observed social information. We create a new bandit-based game environment and provide a demonstration of how this environment can be used to test computational theories about the role of expertise in social learning, allowing for direct comparisons between humans and computational models at scale.

For our computational modeling, we begin by using Thompson sampling, a common probabilistic heuristic for solving multi-armed bandits, and the Upper Confidence Bound (UCB) algorithm, a popular Q-value based multi-armed bandit solver. We chose these solvers due to their effectiveness in modeling human decision making (Gershman, 2018). We then develop hybrid social learning models by training a supervised neural network to predict the teacher's knowledge and integrate these predictions with the parameters of Thompson sampling and UCB. We consider hybrid models that have access to information about both the teacher's action choices and reward outcome information as well as a choice-only variant without access to reward outcomes. We additionally compare to a naïve estimation baseline, which treats social information as additional self-observations. To preview our results, we find that – in contrast to the baseline model – both hybrid social learning models succeed in describing human performance in our task by capturing reward outcomes and decision entropy.

All source code for this work is available on GitHub.[1]

Experiment

Method

We created an engaging multi-armed bandit game, styled as an asteroid mining adventure (see Figure 1). In this game, asteroids serve as the bandit arms, each with its own Bernoulli payoff parameter that remains unknown to the player. Players navigate this space-themed game in a starship using the left and right arrow keys to move and the up arrow or space bar to shoot. They are free to move left or right as often as they like, but opportunities to shoot (pull the lever of the bandit arm) are limited.

Participants 200 adults between the ages of 18 and 45 were recruited via Prolific and were paid $2.50 for participating. Our final dataset included 153 participants. Participants were excluded for colourblindness (N=5) or for refreshing their page during the experiment (N=42), since reloading the page reset the condition assignment.

[1] https://github.com/langcog/social-rl-analysis

Figure 1: A screenshot of our experimental paradigm, showing a multi-armed bandit presented as an asteroid mining adventure. For space, five of a total of 12 asteroids are shown here.

Design The game was structured into three distinct block conditions, each representing a specific game-sequence of play and watch games. The block conditions were: *play-play-play*, *play-watch-play*, and *watch-play-play*, as illustrated in Figure 2. In "play" games, participants actively engaged in the game by moving or shooting, whereas in "watch" games, they observed a teacher – either a novice or an expert – play.

When a player shoots an asteroid, they either receive a reward (1) or not (0), based on the asteroid's underlying Bernoulli payoff parameter. Successful shots, yielding rewards, are visually indicated by a green mark on the asteroid and a corresponding green bar filling up in the on-screen "Mineral Mining Tank", see Figure 1. Conversely, shots that do not result in a reward leave a red mark on the asteroid and add a red bar to the Mineral Mining Tank. A single watch or play game reaches its conclusion once the Mining Tank is completely filled, indicated by the absence of any remaining gray bars. This game design allows players to interact in a multi-armed bandit environment in a visually intuitive and engaging way.

Each of the three games from every block condition was designed to last 24 shot trials. After completing 24 shoot actions, the participant moved to the next game in the game-sequence of the block condition, at which point the mineral mining reward tank was reset to all gray bars, and all shot marks were cleared, indicating the starting state of zero shots and zero rewards for the new game. This framework was applied to all three games of the block condition, ensuring a uniform experience across different sessions.

Two teacher conditions – a novice and an expert – were defined based on the level of proficiency in playing the bandit game. The novice teacher follows an exploratory policy with limited experience in identifying the game's payoffs, leading to a lower expected reward. In contrast, the expert teacher follows a policy developed from a more extensive exploration of the payoffs, resulting in a higher expected total reward.

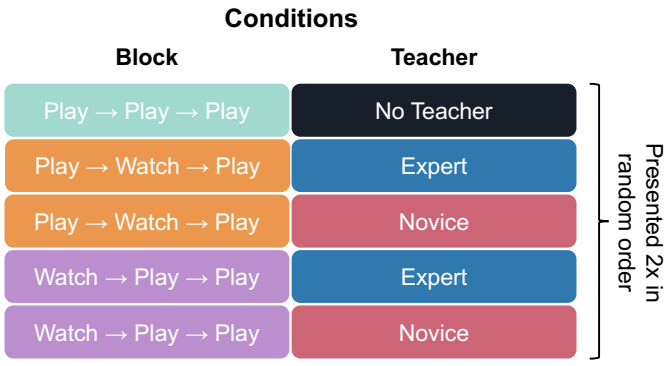

Figure 2: Schematic depiction of experimental conditions.

The teacher is indicated on screen by a separate starship from the participant's starship during watch game. However, the teacher's skill level (novice or expert) is not indicated.

The experiment was designed to explore five unique combinations of block and teacher conditions as shown in Figure 2. Critically, every time a pair of block and teacher conditions was initiated, a random set of stored payoffs was selected indicated by the "zone" in the game. Thus, each zone corresponded to a unique set of payoffs that applied to all games within a block condition. During the watch game, the corresponding teacher policy (novice or expert) was randomly selected from the set of stored policies. This approach allowed participants to repeatedly engage with the payoff environment, either by direct participation or by observing a teacher's gameplay.

We generated a sufficiently large set of payoff zones (100) from which we could randomly sample 10 (one for each set of conditions, twice) for each participant. Each payoff zone consisted of a 12-armed bandit, with payoff parameters sampled as follows: 3 arms with payoffs uniformly sampled between 0 and 0.2, 8 arms uniformly sampled between 0.45 and 0.55, and 1 arm (the optimal arm) uniformly sampled between 0.75 and 1.0. For each of the 100 bandits, we generated a set of 20 novice and expert teachers by running Thompson sampling. 72 iterations of Thompson sampling generated a sequence of arm choices. The first 24 iterations were selected as the novice teacher for the bandit due to their more exploratory nature. In contrast, the final 24 iterations were selected as the expert teacher due to their more frequent exploitation of higher paying arms. A total of 100 candidate multi-armed bandits were generated for the experiment and stored. For each bandit, 20 corresponding novice and expert policies were generated and stored.

Procedure Participants were introduced to the game with a backstory, setting the stage for their asteroid mining adventure titled "Asteroid Fortune Frontier." They were briefed on their mission aboard Proctor One, their virtual starship, where they would traverse different zones in the galaxy. The narrative established that each asteroid harbored unknown rewards and that the distinct zones influenced the hidden payoffs.

We created an initial set of 10 unique asteroid designs. We then varied them through random rotations and hue adjustments, creating a diverse pool of 80 asteroids. In each zone of the game, 12 of these asteroids were randomly selected (one for each of the 12 bandit arms), enhancing the game's variety and unpredictability. Participants were told that successfully mining asteroids would accumulate celestial wealth in their Mining Tank.

Participants were informed about the possibility of observing other starships in action, emphasizing the importance of attention during these observation phases to potentially aid their own gameplay. Prior to diving into the main part the experiment, participants engaged in a practice run to familiarize themselves with the game mechanics. They then proceeded through all 5 condition pairs, each twice, in random order. Each condition pair was presented as a unique zone, labeled as Zone 1, Zone 2, etc. on the screen. Participants were not provided with specific details about the block or teacher condition, or the underlying payoffs in each zone.

Results

Average total reward for each condition and block are shown in Figure 3. All participants improved their total reward over the course of each block, but total reward varied by condition. To quantify variation in learning outcomes, and in particular, how adding watch games impacts total reward, we created a linear mixed effects model predicting mean total reward on the second "play" block for each condition (see Table 1). We chose this block because it allowed us to compare the effects of social learning with different teachers to the same amount of direct exploration without any social learning input.

Full model results are shown in Table 1. Both social learning conditions with an expert teacher yielded significant increases in total reward over pure self-exploration. To explore the significance of the contrast of novice teacher conditions with the self-exploration baseline, we re-parameterized the model so that novice was the reference level. The watch-play-play condition with a novice teacher showed a small but significant increase over the play-play-play baseline ($p < 0.001$) while the play-watch-play condition with a novice teacher was not significantly different ($p = 0.251$). There was no significant difference between the play-watch-play and watch-play-play conditions ($\beta = -0.44, p = 0.177$).

Social Learning Models

We modeled the participants' arm choices and rewards using Thompson sampling-based simulation. We first describe each model and then how we use them to simulate data from our experiment.

Models

To model how subjects integrated social information from the watch games with their own prior knowledge, we used models based on Thompson sampling and UCB. Each Thompson sampler is parameterized through a set of Beta-binomial conjugate distributions that represent the distribution over reward

Table 1: Mixed-Effects Model Results

	Estimate	Std. Error	t-value	df	p-value
(Intercept) **block** : [play-play-play]	15.63	0.27	58.99	1479.00	<0.001
block : [play-watch-play], **teacher** : expert	2.37	0.33	7.11	1479.00	<0.001
block : [watch-play-play], **teacher** : expert	1.92	0.32	6.01	1479.00	<0.001
block : [watch-play-play], **teacher** : novice	-2.49	0.36	-6.98	1479.00	<0.001
block : [play-watch-play], **teacher** : novice	-0.57	0.49	-1.15	1479.00	0.251

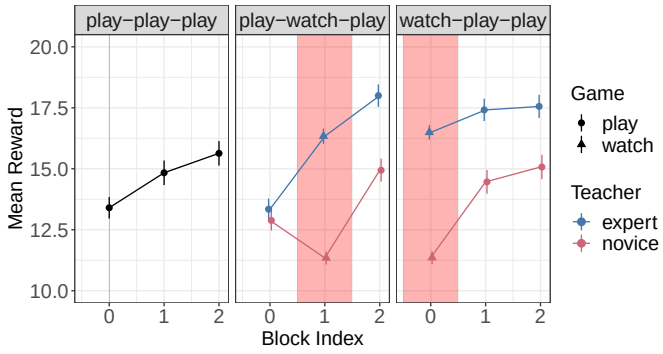

Figure 3: Mean total reward for human participants for each game, indicated by the block condition index, on all pairs of block and teacher conditions. Watch scores, overlaid with a pink hue, represent mean total reward for expert or novice teacher demonstrations. Error bars represent 95% CI.

probabilities in each arm i of the bandit via a pair of Beta parameters, α_i and β_i. Each UCB sampler is parameterized through a set of counts for the number of times each bandit arm has been pulled N_i and the total reward r_i that has been generated from each arm i from which Q-values and the number of time steps can be computed to run UCB.

In Thompson sampling each pair of parameters encodes the number of times an arm has been pulled, with α_i referring the number of times a reward was generated for arm i, and β_i referring to the number of times there was no reward outcome. Thus, for additional arm pulls, updating the parameters only requires knowing the total number of trials with the number of times a reward was generated (for each arm). These parameters are estimated from watch games, denoted α_i^{watch} and β_i^{watch} for each arm i, and then integrated with prior knowledge via an update rule in 1. We can update UCB parameters in the samer manner as shown in 2.

$$\alpha_i \leftarrow \alpha_i + \alpha_i^{watch} \qquad \beta_i \leftarrow \beta_i + \beta_i^{watch} \tag{1}$$

$$N_i \leftarrow N_i + N_i^{watch} \qquad r_i \leftarrow r_i + r_i^{watch} \tag{2}$$

If there is no previous experience (play or watch), then all parameters are 0 prior to integration. Thus integration is accomplished by an additive combination of the Thompson or UCB parameters from prior knowledge and those estimated from watch games. The estimation methods we present next are methods for estimating the watch parameters. We used the same additive integration method regardless of how we estimated the parameters themselves.

Naive Estimation Model In this initial model, we treat observed games as additional self-play observations. We estimate the Beta parameters by following same update rule in Thompson sampling directly from the observed actions and reward outcomes. For UCB we simply count number of arms pulls and rewards for each arm.

Hybrid Choice-Only Model Our next model attempted estimating expertise and using it as a basis for information integration. We trained a supervised neural network to predict ground truth Thompson parameters. We generated a separate training dataset of 12 armed bandits with random payoffs to correspond to the 12 asteroids in the asteroid mining game. This dataset consisted of 10,000 generated bandit problems with Thompson sampling repeated 10 times on each bandit (each for 72 iterations). This generated a training dataset of 100,000 Thompson action sequences, each of length 72, along with the resulting Thompson parameters for each step of every sequence to serve as training targets. We also generated a validation set, using 1000 additional bandit problems, for a total of 10,000 Thompson and UCB action sequences of length 72. Using the training set we trained a Multi-layer Perceptron (MLP), to predict the underlying α_is and β_is corresponding to the end of the sequence of the observations from the Thompson sampler simulations. In other words, given an observation window of 24 actions sequences, each MLP is supervised to predict the Thompson parameters corresponding to the end of that sequence despite not observing actions preceding the observation window. We refer to this model – which does not have access to the reward information – as the "hybrid choice-only estimator." Since the α_is and β_is can be equivalently transformed to N_i's and r_i's and vice versa, we additionally used the same Thompson based hybrid choice-only model for a UCB hybrid choice-only parameter estimator.

Hybrid Choice-Reward Model In order to probe the impact that observing rewards has on the parameter estimation and subsequent game play simulation, we additionally fit a model using predictive estimation but with both choice and reward observation. To do this we trained the same parameter estimator but added reward information from the same observation window to the input during the parameter prediction step. We call this model "hybrid choice-reward estimator". Similar to the hybrid choice-only model, we can use the the Thompson hybrid choice-reward estimator to define a UCB

Table 2: Human-Model Correlations (r) and RMSE

Criteria	Model	Condition Level			Participant Level		
		r	p	RMSE	r	p	RMSE
Reward	Thompson Naive Estimator	0.703	0.023	1.648	0.191	< 0.001	5.201
	Thompson Hybrid Choice-Only	0.974	< 0.001	1.028	0.285	< 0.001	**5.017**
	Thompson Hybrid Choice-Reward	**0.976**	< 0.001	**0.869**	0.269	< 0.001	5.063
	UCB Naive Estimator	0.617	0.057	2.592	0.219	< 0.001	5.552
	UCB Hybrid Choice-Only	0.895	< 0.001	1.864	0.296	< 0.001	5.407
	UCB Hybrid Choice-Reward	0.90	< 0.001	2.042	**0.316**	< 0.001	5.405
Entropy	Thompson Naive Estimator	0.712	0.021	5.703	0.138	< 0.001	8.026
	Thompson Hybrid Choice-Only	**0.930**	< 0.001	5.387	0.153	< 0.001	7.835
	Thompson Hybrid Choice-Reward	0.917	< 0.001	5.221	0.161	< 0.001	7.732
	UCB Naive Estimator	0.556	0.095	4.968	0.133	< 0.001	7.535
	UCB Hybrid Choice-Only	0.789	0.007	4.415	0.150	< 0.001	7.274
	UCB Hybrid Choice-Reward	0.785	0.007	**4.185**	**0.163**	< 0.001	**7.001**

hybrid choice-reward parameter estimator as well.

Simulations

To model human data, we used the same bandits, payoffs, and teacher demonstrations with reward outcomes, shown to participants. We simulated results for every block and teacher condition completed by a participant.

The play-play-play block condition did not require hybrid estimation since there were no watch games. Instead for this block condition we used standard Thompson sampling and UCB respectively.

The play-watch-play condition required first a standard Thompson sampling and UCB run respectively for 24 steps for the 1st play game, resulting in Thompson and UCB parameters for each arm, followed by a watch game simulated by running Thompson-based and UCB-based naive estimation, hybrid choice-only estimation, and hybrid choice-reward estimation. After completion of the watch game, the Thompson and UCB parameters from the first play game and the watch game were integrated using the update rules from equations 1 and 2. The second play game was simulated with Thompson and UCB sampling by first initializing each model's respective parameters with the resulting integrated parameters. The simulation for this block condition resulted in two sequences of 24 play steps of 24 arm pulls and corresponding rewards for the first and second play for each of the three estimation methods based on Thompson sampling and UCB.

The watch-play-play condition started by directly estimating Thompson and UCB parameters of the teacher foe the watch game using each of the three Thompson and UCB estimators. These parameters were then used to initialize three separate Thompson and UCB samplers (one for each corresponding estimator), each of which subsequently simulated 24 arm pulls for the first play game. The second play game followed suit but initializing Thompson and UCB with the parameters resulting from the end of 1st play game and subsequently simulated another 24 arm pulls.

Comparing models to human performance

We compared models to human performance by computing Pearson correlation and root mean squared error (RMSE) of the total reward and entropy over arm choices between humans and each of the three models (naive estimation, hybrid choice-only estimation and hybrid choice-reward estimation) for both Thompson sampling and UCB on the play games. We computed both *condition level* comparisons and *participant level* comparisons. Condition level comparisons are between mean values, of total reward and entropy respectively, across pairs of block and teacher conditions. Participant level comparisons are direct human to model comparisons, of total reward and entropy receptively, for individual games. All comparisons between human and model performance are parameter-free – no parameters of the models were fit to the human data.

Each comparison used three weight initialization seeds for the MLPs in the hybrid models. Furthermore, each of the models used three unique payoff seeds for the multi-armed bandit during simulation over which we computed average reward across seeds with 95% confidence intervals.

We report the comparison results in Table 2. The Thompson hybrid choice-reward model showed the best condition-level correlation and lowest RMSE for mean reward. Thompson hybrid choice-only showed the best entropy correlation with the Thompson hybrid choice-reward showing similar performance. The RMSE values for entropy however were better performed by the UCB hybrid choice-reward model. For participant level, reward and entropy correlations were smaller as expected and best performed by UCB hybrid choice-reward.

We show our best-fitting model, the Thompson hybrid choice-reward, alongside other Thompson variants in Figure 4. This figure displays the average total rewards for the first and second play games in each block condition for par-

ticipants and the Thompson models for different teacher conditions. Thompson sampling, showed an increase in total rewards from the first to the second play game, similar to humans, but with a slightly larger increase than participants. In the play-watch-play and watch-play-play conditions, consistent differences emerged between naive and hybrid estimation. Naive estimation increased in reward between the first and second play. However, the relative increases in reward had significant differences compared to humans. Hybrid estimation, on the other hand, strongly resembled participants in the play-watch-play condition with sharper reward increases from with an expert teacher, as well in the watch-play-play condition with better starting reward in the first play after watching the expert teacher.

General Discussion

This study aimed to provide an environment for studying social learning within the context of multi-armed bandits. We used this multi-armed bandit game to test two models of social learning, naïve estimation – treating social information as additional actions and outcomes – and hybrid estimation – in which neural networks were used to infer parameters estimating experience and expertise from observations. Our human experiment found that participants achieved higher total rewards when they observe an expert teacher compared to a novice teacher or no teacher. The hybrid models performed quite well, especially the Thompson based models, at matching participant data even without any specific fitting to the human data, especially compared with a baseline model that treated social observations the same as self-play.

Interestingly, between our two hybrid social learning models for both Thompson and UCB, the choice-only variant performed nearly as well as the choice-reward variant when comparing mean reward and even slightly better correlated when comparing entropy from bandit arm choices. Hybrid choice-only models were able to estimate the teacher's Thompson parameters from choice data alone with reasonably low error, suggesting that observing reward outcomes is not always necessary for estimating experience level and expertise. This finding raises an interesting question about human social learning: do humans keep track of teacher payoffs, or do they instead care more about teachers' actions? Perhaps actions themselves are enough for social learners to infer expertise.

A notable limitation of this work is the choice to train MLPs to predict Beta parameters, an internal parameter of the Thompson sampler. To the extent they capture internal belief states, these parameters would not be observable, which challenges the claim that integration with Thompson parameter estimation somehow replicates the cognitive process of human social learning. Despite this challenge, predictive estimation of Beta parameters showcases the the capability to precisely estimate experience level and how this process might be incorporated with a learner's own prior knowledge. Thus, this work underscores a promising direction for future research based on the premise that humans employ some form

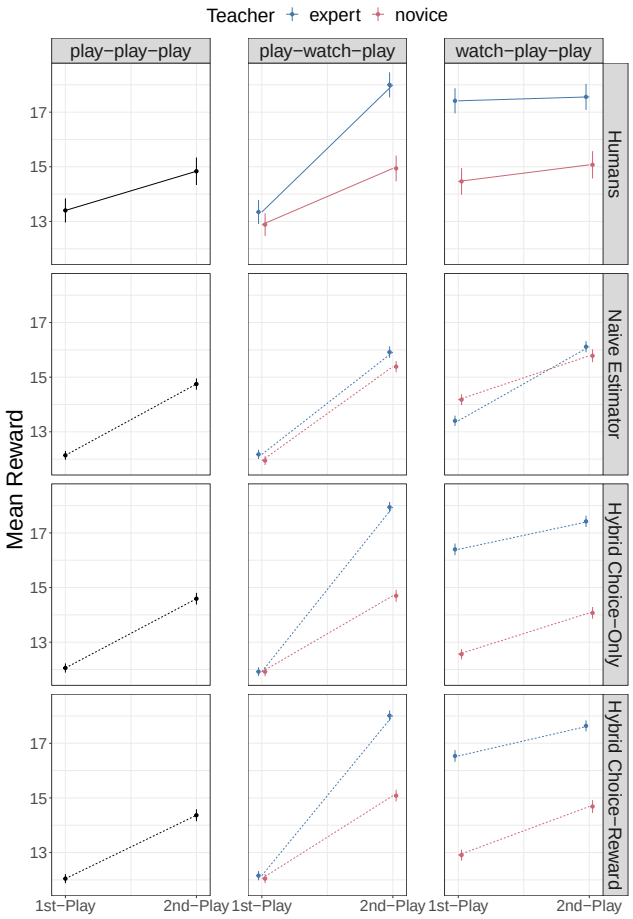

Figure 4: Mean rewards are shown for first and second play games on each block condition. First row shows human rewards and subsequent rows show Thompson model rewards. Mean rewards for play-play-play condition are displayed in black indicating no teacher condition. Blue [red] points indicate results with an expert [novice] teacher. Error bars represent 95% CIs.

of estimation to approximate teacher experience during social learning to infer beyond the observations they see.

This study's scope was deliberately confined to the realm of multi-armed bandits, which inherently presents a limitation. It is evident that social learning transcends the framework of a bandit game characterized by intermittent observation cycles. Social learning dynamics are considerably more complex and dynamic than what our current setup provides. Nonetheless, this constraint has been instrumental in facilitating the comparison of humans to models. Moving forward, we plan to enhance the complexity of our experimental environment. This includes the possibility of contextual bandits and improving the dynamics between observations and actions. We hope that this environment will allow future research to capture some of the intricate dynamics of human social learning.

References

Adrian, J. A., Siddharth, S., Baquar, S. Z. A., Jung, T.-P., & Deák, G. (2019). Decision-making in a social multi-armed bandit task: Behavior, electrophysiology and pupillometry. *arXiv preprint arXiv:1905.07474*.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Behrens, T. E., Hunt, L. T., & Rushworth, M. F. (2009). The computation of social behavior. *science*, *324*(5931), 1160–1164.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249.

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, *80*(6), 1558–1571.

Burtini, G., Loeppky, J., & Lawrence, R. (2015). A survey of online experiment design with the stochastic multi-armed bandit. *arXiv preprint arXiv:1510.00757*.

Charpentier, C. J., & O'Doherty, J. P. (2018). The application of computational models to social neuroscience: promises and pitfalls. *Social neuroscience*, *13*(6), 637–647.

Duranti, A., Ochs, E., & Schieffelin, B. B. (2014). *The handbook of language socialization*. John Wiley & Sons.

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

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Giraldeau, L.-A., Caraco, T., & Valone, T. J. (1994). Social foraging: individual learning and cultural transmission of innovations. *Behavioral Ecology*, *5*(1), 35–43.

Gray, R. C., Zhu, J., Arigo, D., Forman, E., & Ontañón, S. (2020). Player modeling via multi-armed bandits. In *Proceedings of the 15th international conference on the foundations of digital games* (pp. 1–8).

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

Kuleshov, V., & Precup, D. (2014). Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*.

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

Molleman, L., Kurvers, R. H., & van den Bos, W. (2019). Unleashing the beast: A brief measure of human social information use. *Evolution and Human Behavior*, *40*(5), 492–499.

Muthukrishna, M., Morgan, T. J., & Henrich, J. (2016). The when and who of social learning and conformist transmission. *Evolution and Human Behavior*, *37*(1), 10–20.

Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Reviews Neuroscience*, *21*(4), 197–212.

Rosenberg, D., & Vieille, N. (2019). On the efficiency of social learning. *Econometrica*, *87*(6), 2141–2168.

Schulz, E., Franklin, N. T., & Gershman, S. J. (2020). Finding structure in multi-armed bandits. *Cognitive psychology*, *119*, 101261.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2015). Learning and decisions in contextual multi-armed bandit tasks. In *Cogsci.*

Stojic, H., Analytis, P. P., & Speekenbrink, M. (2015). Human behavior in contextual multi-armed bandit problems. In *Cogsci.*

Vélez, N., & Gweon, H. (2021). Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current opinion in behavioral sciences*, *38*, 110–115.