Social Learning via Bayesian Inverse Reinforcement Learning: Learning from and about a Learner

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

Anurag Dutt
Department of Computer Science, Stony Brook University, NY 11794, USA

Christian C. Luhmann
Department Psychology, Stony Brook University, NY 11794, USA

Abstract

What does a social learner learn? Research has explored imitation-based social learning strategies as well as inverse reinforcement algorithms that estimate others' true reward function. In the current study, we propose that social learning may be more elaborate and develop a model of social learning using Bayesian inference that seeks to understand both the task an observed demonstrator is performing and the demonstrator itself. Using simulations, we show that the model is able to learn about the demonstrator when provided with full and partial information. We strengthen this point by asking the model to make inferences about missing choice and reward information. Last, we show that the model is able to represent one set of beliefs about the environment while attributing a distinct set of beliefs to the demonstrator. Thus, we move away from simple models of social learning, investigating inference-making as a core mechanism of social learning.

Keywords: social learning; (inverse) reinforcement learning; Bayesian learner

The act of learning based on the actions of 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 identified as a distinct mechanism that has the potential to offer advantages over individual learning (Henrich & McElreath, 2003; Boyd, Richerson, & Henrich, 2011). Compared to individual learning, social learning is generally considered to be less costly. For example, social learning can speed up the learning process and can allow social learners to avoid severe, potentially fatal, outcomes (Rendell et al., 2011; Boyd et al., 2011). The adaptive potential of social learning has been studied and discussed in a variety of contexts ranging from child development (for reviews see Harris, 2012; Koenig & Sabbagh, 2013) to cultural evolution (e.g., McElreath et al., 2005), measuring individual- (e.g., Najar, Bonnet, Bahrami, & Palminteri, 2020; Toyokawa, Saito, & Kameda, 2017; Toyokawa, Whalen, & Laland, 2019) and group-level outcomes (e.g., Boyd & Richerson, 2009; Rendell et al., 2011). This widespread applicability has led to comprehensive investigations across various disciplines, including evolutionary anthropology and biology, behavioral economics, computer science, psychology, and sociology, highlighting its relevance in numerous fields. This work assumes that the goal of social learning is to better understand one's environment or a given task. In the current paper, we instead propose and evaluate a model of an observational social learner that seeks to use another agent's behavior to understand an environment as well as understand the agent itself. Before doing so, we first review past work on social learning.

What is Social Learning

Social Learning Strategies

Many models of social learning are heuristic in nature. For example, the literature on social learning often discusses the concept of social learning *strategies* (Laland, 2004; McElreath et al., 2005; Gigerenzer & Todd, 1999; Rendell et al., 2010). These social learning strategies determine *what* behavior to copy *when* from *whom*. Consequently, these learning strategies simplify social learning into an act of imitation, proposing a pivotal role for the logic of selecting an appropriate social learning strategy.

Formal Models of Social Learning

There has been considerable research about what social learning strategies *should* be used and what social learning strategies *are actually* used. To determine what specific social learning strategy people should and do employ, strategies have been translated into a broad range of formal, largely heuristic, models often based on a foundation of imitation.

The seminal work of Rogers (1988) kicked of an extensive debate about the adaptability of social learning, demonstrating that social learning does not always enhance a group's fitness—a finding later called Roger's paradox. This finding spurred extensive theoretical research into identifying which social learning strategies increase human fitness and which do not. In decades of subsequent research, researchers such as Boyd and Richerson have developed a range of mathematical models that evaluate the adaptability of social learning—defined as copying behavior—and its impact on cultural evolution (e.g., Boyd & Richerson, 1985, 1988).

Likewise, social learning strategies have been much studied in a behavioral context. For example, McElreath et al. (2005) explored whether people copy the behavior of one other person (linear imitation), stick to their selected choice when another person previously made the same choice (confirmation), or whether people adopt the behavior exhibited by a majority of others (conformity), all while weighing how much impact social information has in contrast to one's own information. Najar et al. (2020) investigated three different ways imitation could be included in a reinforcement learning

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.*

model. Typically, reinforcement learning models include at least two components: an update function representing learning and a choice function. Najar et al. concluded that social information augmented the perceived value of options during learning (a value shaping model) rather than directly augmenting the choice selection process or via a standard inverse reinforcement learning (IRL) process (IRL is discussed further below). In these behavioral studies social learners do not have access to others' rewards but do have access to others' choices. However, Nedic, Tomlin, Holmes, Prentice, and Cohen (2011) explored situations in which either rewards, choices or some combination of both were missing. They developed heuristic models tailored to each situation and used weights in the choice function to bias action selection.

More recently, inference abilities have been proposed as a potential social learning mechanism. Gweon (2021) outlined how inference-making allows to interpret and learn from evidence generated by others through a process they call inferential social learning. Vélez and Gweon (2021) further support the idea that social learning can be formalized as a probabilistic inference process and call to combine RL and Bayesian approaches. Hawkins et al. (2023) developed a mathematical, yet still heuristic-based model that interprets choice patterns in relation to reward magnitude. Specifically, repeated choices (exploitation) imply high rewards and varied choices (exploration) imply low rewards. Hawkins et al. (2023) argue that humans rely on the just described inference process to learn from their partner's choice information.

Inverse Reinforcement Learning (IRL)

Machine learning researchers have also studied algorithms designed to learn from the experiences of others. These algorithms, often referred to as inverse reinforcement learning (Arora & Doshi, 2021), assume that observers are seeking to estimate the agent's reward function; the relationship between the reinforcement provided by the environment and actions taken by an agent (and the states in which those actions are taken in the case of a Markov decision process). Other than the to-be-inferred reward function, IRL algorithms are conventionally assumed to have omniscient access to all other relevant information. This includes task parameters such as the transition matrix describing how the agent's actions in state, S_t, result in a transition to a new state, S_{t+1} as well as the policy that is optimal for the task. Having access to the optimal policy is equivalent to assuming that the agent is optimal and that the algorithm has access to the agent's policy (the mapping from states to actions).

Early work (Ng, Russell, et al., 2000) focused on the core problem of estimating the reward function while avoiding trivial solutions. Since then, work has largely focused on relaxing the assumptions described above. For example, algorithms that use behavioral demonstrations of an optimal agent instead of the optimal policy (Abbeel & Ng, 2004). Others have loosened the assumption that the observed agent is behaving optimally (Jacq, Geist, Paiva, & Pietquin, 2019; Ramponi, Drappo, & Restelli, 2020). Research on IRL has also considered scenarios involving even greater uncertainty, such as those involving partially observable Markov decision processes (POMDPs), settings in which the agent cannot know its current state with certainty (Choi & Kim, 2011; Djeumou, Cubuktepe, Lennon, & Topcu, 2022). Other work has investigating learning when only given partial access to the agent's behavior, sometimes referred to as "occlusion". Bogert and Doshi (2018) acknowledge the realism and the utility of relaxing the full-information assumptions, studying such scenarios in a multi-robot application.

IRL researchers have also considered situations involving humans. Some have investigated scenarios in which an IRL algorithm is tasked with learned from a human rather than from an optimal agent (Hadfield-Menell, Russell, Abbeel, & Dragan, 2016; Pan et al., 2018), developing algorithms that fall into the emerging field of "socially-aware artificial intelligence" (Krishna, Lee, Fei-Fei, & Bernstein, 2022; Lukowicz, Pentland, & Ferscha, 2011). These approaches are far more likely to consider interactive dynamics between the observer and learner rather than a simple one-way observational learning. Other work has investigated scenarios in which it is a human observes the behavior of a computational agent (e.g., a robot) and must use this behavior to learn a task (Lee, Admoni, & Simmons, 2022a, 2022b).

Bridging the gap between technical approaches and psychological theory, it has been argued that IRL can be used to formalize what has been termed "Theory of Mind" (Premack & Woodruff, 1978), the ability to infer and reason about other peoples' hidden mental states (Jara-Ettinger, 2019). Specifically, IRL could be used to infer mental states, and RL could be used to predict other's actions.

Summary

To briefly summarize, there have been a variety of approaches to social learning, though there are some notable themes that pervade this work. Social learning research in the social sciences focused on strategies that are heuristic in nature and essentially assumes that the core mechanism of social learning is the mimicry of others' behavior. The research objective is then to characterize how such mimicry is deployed. IRL, in contrast, takes a strongly principled approach, but in doing so requires tremendous amount of precise knowledge to provide any guarantees. Furthermore, the objectives of IRL are quite modest; algorithms are simply trying to estimate agents' reward function. Finally, the formal models of social learning developed within the social sciences are still based on heuristic strategies and largely adopt the modest objectives of IRL (e.g., learn about unobserved rewards while having access to nearly all other relevant information).

The Current Study: An Optimal Model of Social Learning

In the current study, we move away from a simple model of social learning, beyond mere imitation, and beyond the objective of learning the true reward function. Our approach

expands upon earlier work that has either taken a heuristic-approach using social learning strategies or focused on different problem formulations such as those typically assumed in IRL. We develop a Bayesian model of social learning that seeks to understand both the task an observed agent is performing and the agent itself.

In what follows, we first describe the context in which this social learner operates, outlining the environment and task in the problem statement. We then formally define a Bayesian inference model. Then, the capabilities of the model are thoroughly tested using a variety of different scenarios. We demonstrate that the model is able to learn from other agents as it develops a reasonable representation of the learning environment and the agents themselves. Moreover, our evaluation extends to scenarios involving missing information, highlighting the model's ability to "fill in" such gaps via inference. Finally, we illustrate how our model can learn about the environment and the other agent, even when the beliefs of the two diverge (cf., IRL). These examinations underscore the inferential power of our model, ultimately prompting questions as to whether human social learning may be a more complex process than previously assumed.

Problem Statement

We use a classical two-armed bandit environment. An agent is asked to repeatedly make a choice, c, between two options, A and B. When selected, each arm generates a reward, r, that is drawn from a normal distribution $\mathcal{N}(\mu_c,\, \sigma^2)$. The agent's goal is to maximize their cumulative reward over a sequence of T trials requiring the agent to learn about the underlying reward function. We assume that these choices and rewards are observed by a second agent. We refer to the agent making choices and receiving rewards as the *demonstrator* and the agent observing the demonstrator as the *observer*. Consistent with many real-world situations, the observer may have incomplete information about the demonstrator's choices and rewards. For example, on a given trial, the observer may have access to the demonstrator's choice but not the associated reward, or to the reward but not the choice. Each of these cases requires a specific treatment.

Our Approach: Social Bayesian RL Model

Using the set of choices, c_t, and the set of observed rewards, r_t, we ultimately wish to infer two sets of quantities. The first is the expected reward of each arm, μ, which is similar to the conventional objective of IRL. The second, however, is the set of parameter values, α and β, that underlie the demonstrator's behavior. These parameter values, combined with the observed information about choices and rewards, allows the observer to infer the demonstrator's beliefs at any given point in the trial sequence.

Inference is characterized as in Equation 1. The payoff function described above specifies $p(r_t|\mu,c_t)$. We assume that the demonstrator begins the task with no information about the reward function (i.e., μ) at all, and thus initial Q-values, Q_1, are assumed to be zero. This contrasts with the standard IRL approach of assuming an optimal demonstrator.

$$\begin{aligned} p(\alpha,\beta,\mu|C,R) \propto & \prod_{t=2}^{T} p(r_t|\mu,c_t)p(c_t|Q_t)p(Q_t|r_{t-1},c_{t-1},Q_{t-1},\alpha,\beta) \\ & \times p(r_1|\mu,c_1)p(c_1|Q_1)) \\ & \times p(\alpha,\beta,\mu) \end{aligned} \tag{1}$$

We assume that the demonstrator is a simple Q-learner. The demonstrator's choice function, which specifies $p(c_t|Q_t)$, is assumed to be a standard softmax

$$p(c_A|Q) = 1 - p(c_B|Q) = \frac{e^{\beta Q_A}}{e^{\beta Q_A} + e^{\beta Q_B}} \tag{2}$$

which depends on a determinism parameter, β. The update rule, which specifies $p(Q_t|r_{t-1},c_{t-1},Q_{t-1},\alpha,\beta)$, is the standard Q-learning update

$$Q_{c,t} = \alpha(r_{t-1} - Q_{c,t-1}) \tag{3}$$

which depends on a learning rate, α. We consider the situation in which the observer has access to some choices and rewards, but others, $\tilde{c_t}$ and $\tilde{r_t}$, are unobserved. In this case, we simply marginalize over the unobserved quantities. Inferred values of α, β, and μ permit the model to make online predictions and inferences about individual unobserved choices or rewards. On a trial in which the demonstrator's choice is observed, the reward (whether unobserved or yet-to-be-observed) can be predicted as $p(r_t|\mu,c_t)$. Choices (whether unobserved or yet-to-be-observed) can be predicted as $p(c_t|Q_t)$. On a trial in which the reward is observed but the choice is unobserved, the unobserved choice can be inferred conditional on the observed reward

$$p(c_{A,t}|r_t) = \frac{p(r_t|c_{A,t})p(c_{A,t})}{p(r_t|c_{A,t})p(c_{A,t}) + p(r_t|c_{B,t})p(c_{B,t})} \tag{4}$$

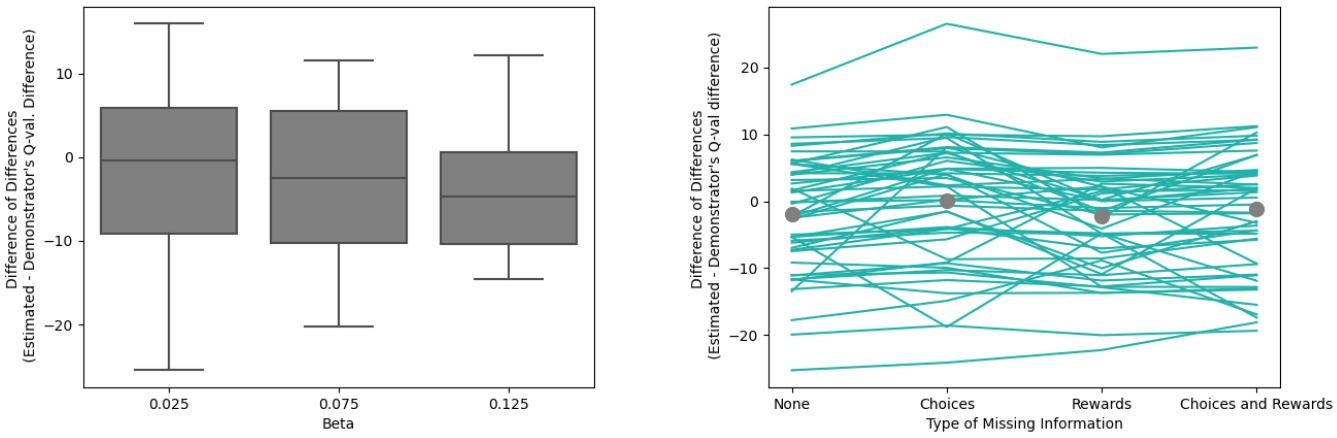
where $p(c_{A,t}) = p(c_{A,t}|Q_{t-1})$. If the reward is unobserved on trial t, but the choice is observed on trial $t+1$, the unobserved reward can be inferred conditional on the immediately succeeding choice

$$p(r_t|c_{t+1}) = \frac{p(c_{A,t+1}|r_t)}{p(c_{A,t+1}|r_t) + p(c_{B,t+1}|r_t)} \tag{5}$$

where $p(c_{A,t+1}|r_t) = p(c_{A,t+1}|Q_{t+1})p(Q_{t+1}|r_t)$.

These inferences were implemented as a probabilistic program, written in PyMC (Abril-Pla et al., 2023). The inferred parameters were given weakly informative priors, largely to confine their values to valid ranges (Eq. 6). $Q_{c=A,t=0}$ and $Q_{c=B,t=0}$ were assumed to be 0.

$$\begin{aligned} \alpha &\sim \mathcal{U}(0,1) \\ \beta &\sim |\mathcal{N}(0,1)| \\ \mu &\sim \mathcal{U}(-10,10) \end{aligned} \tag{6}$$



(a) (b)

Figure 1: (a) Difference between the difference of observer's estimated Q-values of the demonstrator at the end of the trial sequence and the actual difference between demonstrator's Q-values at the end of the trial sequence for three demonstrators ranging from more ($\beta = .025$) to less ($\beta = .125$) stochastic choice behavior, summarizing over 50 choice sequences for each β. (b) Same difference as in (a) for a demonstrator with $\alpha = .1$, and $\beta = .1$, with either no, choice and/or reward information missing. Lines represent 50 trial sequences. Dots represent means.

The posterior distribution (e.g., Eq. 1) was approximated via PyMC's slice sampling MCMC algorithm. In each scenario (or replicate of a scenario), we ran four chains. A total of 1000 samples were used in each chain for tuning (and discarded) and an additional 1000 samples were drawn, for a total of 4000 posterior samples. Sampling diagnostics were inspected manually and were satisfactory in all instances.

Results

As a first step, it is essential to verify that the model forms a reasonable representation of the demonstrator. Here, we specifically investigate whether the model develops precise estimates of the demonstrator's Q-values after observing a sequence of trials (i.e., choices and rewards). Furthermore, we investigate these estimates as a function of the amount of decision noise (i.e., β) reflected in the demonstrator's behavior.

Result 1: The model learns about the demonstrator when given full information

We generated 150 choice sequences using an algorithmic demonstrator characterized by the update and choice function defined in Equations 2 and 3, with a learning rate of $\alpha = .1$. If $Q_{c=A} = -10, Q_{c=B} = 10$ and $\beta = .025$, the demonstrator exhibits more stochastic behavior, choosing the higher- and lower-reward arms with $p = .62$ and $p = .38$ respectively. When $\beta = .075$, the demonstrator chooses the higher- and lower-reward arms with $p = .82$ and $p = .18$ respectively. Lastly, when $\beta = .125$, the demonstrator exhibits more deterministic behavior, choosing the higher- and lower-reward arms with $p = .92$ and $p = .08$ respectively. The task was parameterized such that $\mu_A = -10$, $\mu_B = 10$, and $\sigma = 1$. The sequence consisted of 40 trials.

Figure 1a shows that the observer's estimates of the demonstrator's Q-values are, on average, quite accurate, irrespective of the level of stochasticity in the demonstrator's choices. The difference between the difference of the observer's estimated Q-values of the demonstrator and demonstrator's actual difference of their Q-values is near 0 across all levels of β. The variability in the data is primarily driven by the model's uncertainty about the parameters that govern demonstrators' learning. This uncertainty is influenced by the priors. Overall, we can conclude that the observer's estimates of the demonstrator's beliefs about the arms' values are relatively robust and not significantly affected by increased decision noise on the part of the demonstrator. The model forms accurate beliefs about the demonstrator when given full information.

Result 2: The model learns about the demonstrator when given partial information

Next, we investigated whether the model forms a reasonable representation of the demonstrator even when some information about the demonstrator is missing. To do so, we strategically removed information about what choices the demonstrator made or rewards the demonstrator received.

We generated 50 trial sequences using a similar algorithmic demonstrator, with a learning rate of $\alpha = .1$ and a stochasticity parameter of $\beta = .1$. From these sequences, we randomly selected 10 trials on which information was censored. On these 10 trials, we either censored information about the

Scenario	Trial 1		Trial 2		Trial 3		Trial 4		Trial 5		HDI		Mean
	Arm	Rew.	Arm	Rew.	Arm	Rew.	Arm	Rew.	Arm	Rew.	3%	97%	
1	B	10	A	-10	B	10	A	-10	Infer	10	1.0	1.0	1.0
2	B	10	NA	-10	B	NA	A	-10	Infer	10	1.0	1.0	1.0
3	B	10	A	-10	B	10	A	Infer	B		-9.99	-8.12	-9.19
4	B	10	NA	-10	B	NA	A	Infer	B		-9.97	-7.05	-8.86

Note. Choice and reward sequence for four exemplary scenarios. Scenarios 1 & 2 show model results when inferring demonstrator's choice. Scenarios 3 & 4 show model results when inferring demonstrator's reward. Scenarios 1 & 3 show sequences with complete knowledge about demonstrator's choices and rewards. Scenarios 2 & 4 show sequences with missing information about demonstrator's choices and rewards. "NA" represents missing information. "Infer" shows the quantity to be inferred. Mean and HDIs for choice inferences shows probability that arm B was selected. Mean and HDIs for reward inferences show the grand mean (i.e., potential rewards weighted by their associated probabilities) of the reward value.

Table 1: Scenarios Showcasing Inference Ability

demonstrator's choice, demonstrator's rewards, or we censored choice information on five of the ten trials and censored reward information on the other five trials. In total, this yielded 50 sets of trial sequences each represented by one line in Figure 1b.

Figure 1b shows that the observer's representation of the demonstrator's Q-values was, on average, quite accurate and did not exhibit any obvious relationship to the type of information that was omitted. This indicates that the model is robust, functioning effectively whether it has access to all of the demonstrator's information or only some. Visual inspection of Figure 1bb suggests that accuracy did vary across condition in a small number of sequences. It is possible that these sequences had particularly relevant information censored (e.g., trials early in the sequence). Future work is needed to investigate such possibilities. Overall, however, accuracy was quite uniform and we can conclude that the model successfully forms accurate beliefs about the demonstrator even when confronted with partially censored information.

Result 3: The model makes reasonable inferences when given partial information

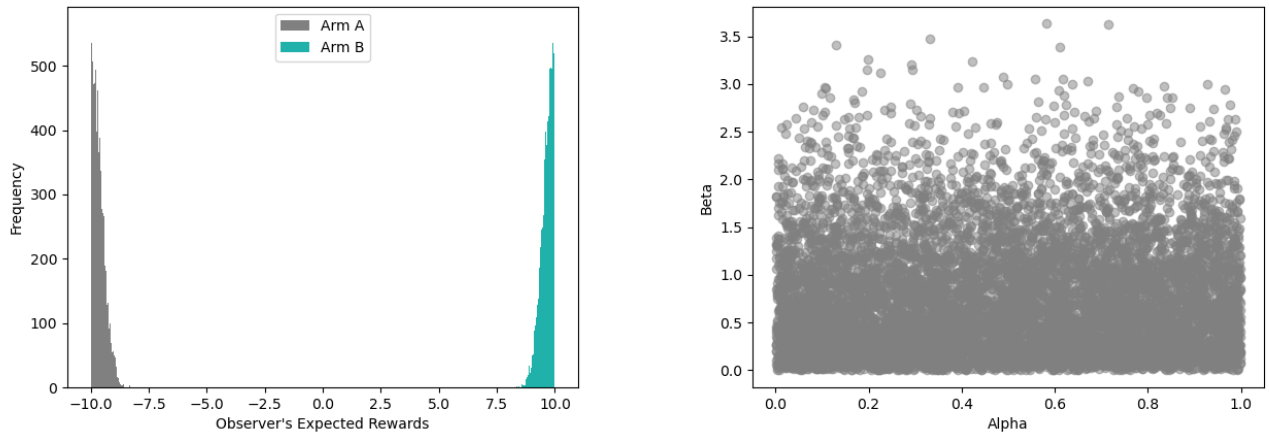
We next investigated whether the model is able to make reasonable inferences about missing information. The inferences of most interest are those in which the model uses observations to draw conclusions about information missing earlier in the trial sequence. For example, if the observer did not observe a demonstrator's choice on trial t, but did observe the reward obtained on trial t, the observer can infer the choice made, (e.g., $p(c_{A,t}|r_t)$). This analysis underscores the model's ability to "fill in" incomplete information.

To illustrate the model's proficiency, we examine four qualitative scenarios detailed in Table 1. Each scenario presents a sequence of five trials involving two arms which yield deterministic rewards of -10 and 10. In the first scenario, a reasonable observer would infer that Arm A is associated with a reward of -10 and Arm B with a reward of 10. Therefore, a straightforward test of the model's capability is to ask which arm was selected on trial five, conditional on the observation of a reward of 10. The model is certain that the demonstrator selected Arm B in this scenario. To increase the complexity, we then tested the model's response to a choice inference in Scenario 2. We use a similar sequence as before but with missing information. Despite the missing information, the model is still certain that the demonstrator selected Arm B.

Scenarios 3 and 4 mirror scenarios 1 and 2, except that instead of a choice inference we have the model generate inferences about unobserved rewards. Hence, instead of the arm on trial five, we now ask about the reward on trial four, conditional the choice on trial five. In a scenario with complete information through the first three trials, the model estimates the missing reward on trial four to be approximately -9.29, with a 94% Highest Density Interval (HDI) suggesting the most credible reward estimates range between -9.99 and -8.12. When faced with a similar query under conditions of missing information (Scenario 4), the model's inference closely aligns with that of the fully informed scenario, offering a mean reward estimate of -9.97 and an HDI spanning from -9.97 to -7.05. Overall, these results confirm that the model is able to make accurate inferences about unobserved choice and rewards. The model can learn even from scenarios involving missing information and can make valid inferences so as to fill in such missing information.

Result 4: The model learns about the environment when the demonstrator is unreasonable

The previous results have shown that the model is able to learn about reasonable demonstrators given either full- and partial-information. In our final investigation, we wished to investigate how the model is able to learn about the environment while simultaneously representing the demonstrator's beliefs (potentially very different beliefs) about the environment. This capability represents a fundamental departure from prior research which has traditionally relied on social learning strategies and IRL. To distinguish the demonstrator's and the observer's understandings of the environment, we in-



(a) (b)

Figure 2: Demonstrator makes 10 alternating choices of arm A and arm B. The associated reward is always -10 for arm A and 10 for arm B. (a) Observer's expected rewards for each arm. (b) Observer's estimation of the demonstrator's parameters.

troduce a scenario with an unreasonable demonstrator.

The demonstrator used here alternates between the two arms despite a clear difference in the rewards each yields. Specifically, the demonstrator first selects Arm B, receives a reward of 10, then selects Arm A and receives a reward of -10, repeated for a total of 10 trials (a reasonable demonstrator would be expected to show a preference for the higher-paying Arm B). If the observer's understanding of the environment exclusively relied on the demonstrator's actions, the observer's understanding would be similarly unreasonable.

Figure 2a shows the model's posterior distribution for each arm's expected value. The true reward for Arm B is 10 and the model's estimates closely aligns with this. Similarly, the true reward for Arm A is -10 and the model's estimates closely aligns with this true reward. This demonstrates that the model is able to build a reasonable representation of the environment despite unreasonable behavior of the demonstrator.

Figure 2b shows the model's posterior distribution of α and β, the beliefs about the nature of the demonstrator. The model tries to make sense of the demonstrator's behavior, but ultimately learns little about either α or β. The demonstrator's "nonsensical" behavior could reflect a very low learning rate, very high decision noise, or some combination of both. Thus, it is unclear what, exactly, the demonstrator believes about the two arms. Despite this, the model learns about the environment in the expected manner.

Discussion

In this paper, we have developed a model of social learning that seeks to understand both the task an observed agent is performing and the agent itself. We demonstrate that this model effectively learns in environments with incomplete information. The model forms own beliefs by inferring missing choice and reward information, constructing expectations and forming a representation of the demonstrator.

Although we have tested a range of scenarios, future research should validate the model across a more diverse set of conditions. For example, a broader range of environments with more arms, or other reward structures should be investigated. Additionally, a broader range of demonstrator parameter values and more variety in how much information is censored could also be investigated. Another area for exploration is the role of initial Q-values. In the current study, we set initial Q-values of each arm to 0. However, one could allow for both arms to be different, and allow for different priors.

The current study allows for a thorough investigation of social learning by providing a comprehensive framework that mirrors the complexities of everyday learning and that integrates different theoretical approaches. Our model acknowledges that in many real-life situations, individuals may not have access to all information about others. Further, the model reflects insights from research on Theory of Mind (e.g., Schenkel, Marlow-O'Connor, Moss, Sweeney, & Pavuluri, 2008; McKinnon & Moscovitch, 2007) in that it forms a representation about what others potentially think about the environment. This feature allows the model to be used in increasingly complex scenarios. For example, in environments in which multiple agents interact, it may be advantageous to not only learn about the environment but also have an idea about what others believe. The model can make predictions about others' actions, which allows for strategic behavior.

Overall, by accurately filling in the gaps of missing information, the model enhances our understanding of learning processes in social contexts. It opens up new possibilities for exploring the complexities inherent in social learning and offers inference-making as one potential social learning mechanism.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on machine learning* (p. 1).

Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., ... others (2023). Pymc: a modern, and comprehensive probabilistic programming framework in python. *PeerJ Computer Science*, *9*, e1516.

Arora, S., & Doshi, P. (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, *297*, 103500.

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

Bogert, K., & Doshi, P. (2018). Multi-robot inverse reinforcement learning under occlusion with estimation of state transitions. *Artificial Intelligence*, *263*, 46–73.

Boyd, R., & Richerson, P. (1988, 01). An evolutionary model of social learning: The effects of spatial and temporal variation. In (p. 29-48).

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. 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.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*(2), 10918–10925.

Choi, J.-D., & Kim, K.-E. (2011). Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, *12*, 691–730.

Djeumou, F., Cubuktepe, M., Lennon, C., & Topcu, U. (2022). Task-guided inverse reinforcement learning under partial information. In *Proceedings of the international conference on automated planning and scheduling* (Vol. 32, pp. 53–61).

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in cognitive sciences*, *25*(10), 896–910.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, *29*.

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

Hawkins, R. D., Berdahl, A. M., Pentland, A. S., Tenenbaum, B, J., Goodman, N. D., & Krafft, P. (2023). Flexible social inference facilitates targeted social learning when rewards are not observable. *Nature Human Behaviour*, *7*(10), 1767–1776.

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, *12*(3), 123–135.

Jacq, A., Geist, M., Paiva, A., & Pietquin, O. (2019). Learning from a learner. In *International conference on machine learning* (pp. 2990–2999).

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

Koenig, M. A., & Sabbagh, M. A. (2013). Selective social learning: new perspectives on learning from others. *Developmental Psychology*, *49*(3), 399.

Krishna, R., Lee, D., Fei-Fei, L., & Bernstein, M. S. (2022). Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, *119*(39), e2115730119.

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

Lee, M. S., Admoni, H., & Simmons, R. (2022a). Counterfactual examples for human inverse reinforcement learning. In *Workshop on explainable agency in artificial intelligence at aaai conference on artificial intelligence.*

Lee, M. S., Admoni, H., & Simmons, R. (2022b). Robot teaching for human inverse reinforcement learning. In *Workshop on robots for learning at acm/ieee international conference on human-robot interaction.*

Lukowicz, P., Pentland, S., & Ferscha, A. (2011). From context awareness to socially aware computing. *IEEE pervasive computing*, *11*(1), 32–41.

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.

McKinnon, M. C., & Moscovitch, M. (2007). Domain-general contributions to social reasoning: Theory of mind and deontic reasoning re-explored. *Cognition*, *102*(2), 179–218.

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., Tomlin, D., Holmes, P., Prentice, D. A., & Cohen, J. D. (2011). A decision task in a social context: Human experiments, models, and analyses of behavioral data. *Proceedings of the IEEE*, *100*(3), 713–733.

Ng, A. Y., Russell, S., et al. (2000). Algorithms for inverse reinforcement learning. In *Icml* (Vol. 1, p. 2).

Pan, X., Ohn-Bar, E., Rhinehart, N., Xu, Y., Shen, Y., & Kitani, K. M. (2018). Human-interactive subgoal supervision for efficient inverse reinforcement learning. *arXiv preprint arXiv:1806.08479*.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Ramponi, G., Drappo, G., & Restelli, M. (2020). Inverse reinforcement learning from a gradient-based learner. *Advances in Neural Information Processing Systems*, *33*,

2458–2468.

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.

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

Schenkel, L., Marlow-O'Connor, M., Moss, M., Sweeney, J., & Pavuluri, M. (2008). Theory of mind and social inference in children and adolescents with bipolar disorder. *Psychological Medicine*, *38*(6), 791–800.

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.

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.

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