Effectively Learning from Pedagogical Demonstrations

Mark K Ho (mark_ho@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences, 190 Thayer Street
Providence, RI 02906 USA

Michael L. Littman (mlittman@cs.brown.edu)
Department of Computer Science, Brown University, 115 Waterman Street
Providence, RI 02912 USA

Fiery Cushman (cushman@fas.harvard.edu)
Department of Psychology, 1484 William James Hall, 33 Kirkland St.
Cambridge, MA 02138 USA

Joseph L. Austerweil (austerweil@wisc.edu)
Department of Psychology, University of Wisconsin-Madison, 1202 W Johnson Street
Madison, WI 53706 USA

Abstract

When observing others' behavior, people use *Theory of Mind* to infer unobservable beliefs, desires, and intentions. And when *showing* what activity one is doing, people will modify their behavior in order to facilitate more accurate interpretation and learning by an observer. Here, we present a novel model of how demonstrators act and observers interpret demonstrations corresponding to different levels of recursive social reasoning (i.e. a cognitive hierarchy) grounded in Theory of Mind. Our model can explain how demonstrators show others how to perform a task and makes predictions about how sophisticated observers can reason about communicative intentions. Additionally, we report an experiment that tests (1) how well an observer can learn from demonstrations that were produced with the intent to communicate, and (2) how an observer's *interpretation* of demonstrations influences their judgments.

Keywords: Theory of Mind; Communicative Intent; Cognitive Hierarchy; Reinforcement Learning; Bayesian Pedagogy

Introduction

People often learn by observing others' demonstrations. Consider learning how to tie your shoes. It would be difficult to learn shoe tying through trial-and-error, which is why we usually learn how to do it from others. However, by itself, being in the presence of social others who are adept at tying shoes is insufficient: imagine trying to learn to tie your shoes by only examining finished knots or briefly watching as someone ties their shoes before rushing out the door. That would be difficult. Instead, people often engage in *teaching interactions* in which a demonstrator *intentionally communicates* the structure of a task or skill while an observer intently watches, *aware* of the demonstrator's pedagogical aims. The demonstrator, to better teach, might modify their behavior to better disambiguate a task, while the observer, to properly learn, might interpret actions in light of these teaching goals to draw better inferences. This form of interaction supports learning in a variety of domains, from learning everyday tasks like shoe tying to complex technical skills to nuanced social norms. Understanding the cognitive processes that support this capacity is thus critical for a painting a complete picture of folk pedagogy and cultural learning (Tomasello et al., 2005; Boyd, Richerson, & Henrich, 2011).

We examine learning from demonstration from the perspective of Theory of Mind (Dennett, 1987; Baker, Saxe, & Tenenbaum, 2009) and communication via recursive social reasoning (Sperber & Wilson, 1986; Shafto, Goodman, & Griffiths, 2014). Theory of Mind is the capacity to reason about one's own or others' mental states (such as beliefs, desires, and intentions) and interpret behavior in light of these mental states. Previous work focuses on how observers reason about agents that are simply *doing* activities such as pursuing goals (Gergely, Nádasdy, Csibra, & Bíró, 1995) or interacting with others besides the observer (Heider & Simmel, 1944; Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013). However, people often intentionally teach (Csibra & Gergely, 2009) and demonstrators who are *showing* how to do a task behave in ways that differ systematically from those simply *doing* a task (Ho, Littman, MacGlashan, Cushman, & Austerweil, 2016).

Here, we present a new framework for modeling how people teach by and learn from demonstrations that combines elements of planning (Puterman, 1994; Sutton & Barto, 1998) and cognitive hierarchy (Camerer, Ho, & Chong, 2004). This has several theoretical advantages and can capture new aspects of data originally reported in Ho et al. (2016). We develop a model of *sophisticated* observers who not only reason about another agent doing a task, but also reason differently about a demonstrator's *communicative* versus *non-communicative* goals, thus learning more effectively than a *naïve* observer who is insensitive to this distinction. Finally, we present the results of an experiment in which participants observed the behavior of another agent doing or showing how to do a task, and participants were told either that they were or were not produced with communicative intent. The model shows a correspondence to peoples' judgments, providing further support for this framework for modeling teaching with and learning from demonstration.

Modeling Teaching with and Learning from Demonstration

To model demonstrator behavior and observer inferences, we draw on two approaches. Aspects of Theory of Mind have been modeled as *inverse planning* (Baker et al., 2009). Meanwhile, recursive social reasoning has been modeled as a cognitive hierarchy (Camerer et al., 2004), where inferences and actions result from Bayesian agents modeling one another. This has been applied to domains such as pragmatics (Frank & Goodman, 2012) and strategic games (Wunder, Kaisers, Yaros, & Littman, 2011). Building on these approaches, we introduce a new model of showing as *planning in observer belief space* and a model of *learning from showing*.

Theory of Mind as Inverse Planning

Markov Decision Processes (MDPs) (Puterman, 1994) can be used to model intentional action (that is, *doing* a task) and serve as the generative model for Theory of Mind inference (Baker et al., 2009). A ground MDP, $M_i \in \mathcal{M}$ is a tuple $< \mathcal{S}, \mathcal{A}, T_i, R_i, \gamma >$: a set of ground states $\mathcal{S}$; a set of actions $\mathcal{A}$; a transition function that maps states and actions to distributions over next states, $T_i : \mathcal{S} \times \mathcal{A} \to P(\mathcal{S})$; a reward function that maps state/action/next-state transitions to scalar rewards, $R_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$; and a discount factor $\gamma \in [0,1)$ that captures a preference for earlier rewards or completing a task quickly. Associated with each MDP is an *optimal value function*, $Q_i^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Intuitively, this corresponds to the maximum expected cumulative discounted reward (i.e. the *value*) that an agent could expect to receive when taking an action a from a state s and acting optimally from then on. Formally, it is uniquely determined by the fixed-point of the recursive Bellman equations $Q_i^*(s,a) = \sum_{s'} T_i(s' \mid s,a)\Big[R_i(s,a,s') + \gamma \max_{a'} Q_i^*(s',a')\Big]$, for each state s and action a. Q_i^* represents the value for a perfectly optimal agent. To account for deviations from this, we assume each "doing" agent uses a soft-max policy, which is defined as:

$$\pi_i^{\text{Do}}(a_t \mid s_t) = \frac{\exp\{Q_i^*(s_t,a_t)/\tau^{\text{Do}}\}}{\sum_{a' \in \mathcal{A}(s_t)} \exp\{Q_i^*(s_t,a')/\tau^{\text{Do}}\}}, \qquad (1)$$

where $\tau^{\text{Do}} > 0$ is a temperature parameter.

Given a generative model of a demonstrator's actions as produced from possible desires (i.e. reward functions), environment knowledge (i.e. states and transitions), and approximate rationality (i.e. acting to soft-maximize value), an observing agent can perform Bayesian inference over worlds (i.e. MDPs). Suppose, at time t, an observer has an initial belief over possible MDPs, $b_t^{\text{Obs}}(M_i)$. As they observe a demonstrator take an action and transition to a new state, they will update their beliefs in accordance with Bayesian inference:

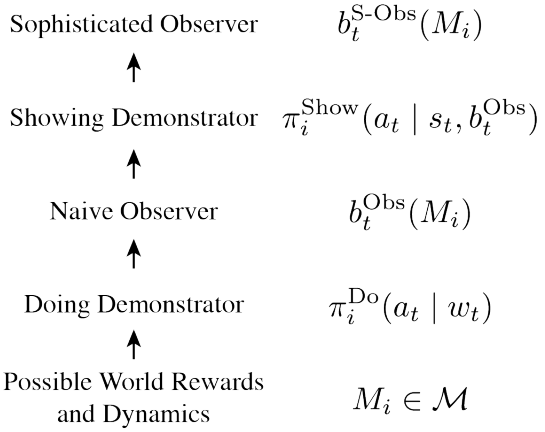

Figure 1: Cognitive hierarchy levels and model notation.

$$\begin{aligned} b_{t+1}^{\text{Obs}}(M_i) &= P(M_i \mid s_t, a_t, s_{t+1}) \\ &\propto P(a_t, s_{t+1} \mid s_t, M_i) P(M_i) \\ &= P(a_t \mid s_t, M_i) P(s_{t+1} \mid s_t, a_t, M_i) P(M_i) \\ &= \pi_i^{\text{Do}}(a_t \mid s_t) T_i(s_{t+1} \mid s_t, a_t) b_t^{\text{Obs}}(M_i). \end{aligned} \qquad (2)$$

That is, at each timestep, an observer's beliefs is updated based on the prior belief from the previous timestep, b_t^{Obs}, the likelihood of the observed state-action transition as given by the optimal policy, π_i^{Do}, and the transition dynamics, T_i, under the MDP M_i. For notational convenience, we define a belief update function $\text{BU}(s_t, a_t, s_{t+1}, b_t^{\text{Obs}})$. The output of this function is b_{t+1}^{Obs}, the observer's belief that the demonstrator is in each MDP given a state-action transition and previous beliefs b_t^{Obs}.

Showing as Planning in Observer Belief Space

An observer can interpret a demonstrator's behavior as *doing* a task using Theory of Mind. But what if the demonstrator is aware that they are being observed and motivated to *show* what task they are performing? Then they may reason not only about how their actions cause transitions and rewards in the ground state-space, $\mathcal{S}$, but also in the observer's belief space, $\Delta_{\mathcal{M}}$. We formulate a *showing* demonstrator, then, as representing an Observer Belief MDP (OBMDP), M_i^{Show}, defined by the tuple $< \mathcal{B}, \mathcal{A}, T_i^{\text{Show}}, R_i^{\text{Show}}, \gamma^{\text{Show}} >$: a joint belief-ground state-space, $\mathcal{B} = \mathcal{S} \times \Delta_{\mathcal{M}}$; the original ground actions, $\mathcal{A}$; a belief-ground state transition function, $T_i^{\text{Show}} : \mathcal{B} \times \mathcal{A} \to \mathcal{B}$; a showing reward function, $R_i^{\text{Show}} : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$; and a showing discount rate, $\gamma^{\text{Show}} \in [0,1)$. (Note from now on we will refer to the doing discount as γ^{Do} to explicitly distinguish it from γ^{Show}.)

We draw attention to two key components of our model of *showing as planning in observer belief space*: the showing reward function, R_i^{Show}, and the showing transition function, T_i^{Show}. They are both influenced by the environment and observer's belief state. When cooperatively showing what they

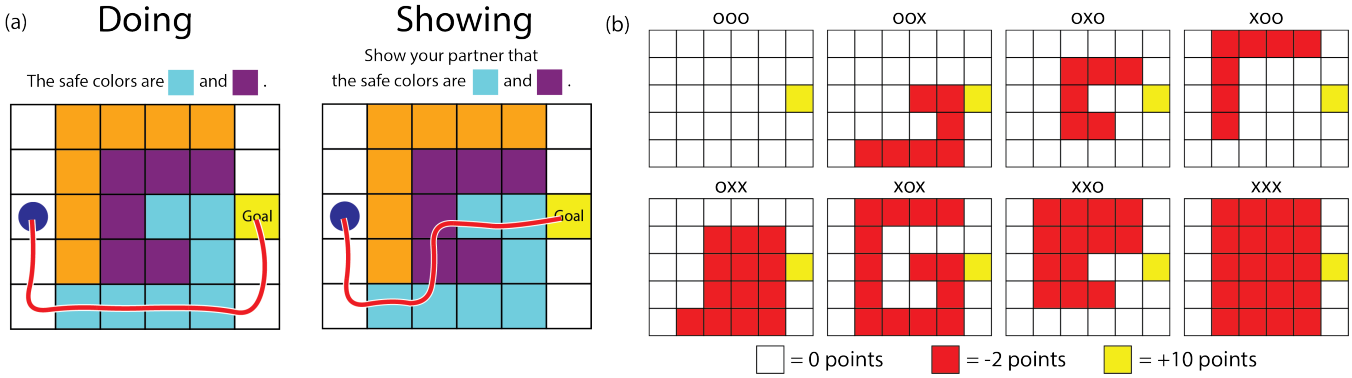

Figure 2: (a) Example of behavior in doing versus showing demonstration conditions from Ho et al. (2016) and the instructions for each condition. (b) Space of eight possible reward functions with different assignments of safe (0 points) and dangerous (-2 points) to each of the three colors. Labels like "ooo" are used later in this paper to reference specific reward functions. Note that participants never directly saw the reward function and inferred them from instructions on a trial.

are doing or how to do a task, a demonstrator *wants* the observer to increase their belief in the true ground MDP, M_i. However, they are still constrained by the rewards in the environment, determined by R_i. That is, they must plan actions with both communicative and non-communicative rewards in mind. Thus, we formulate R_i^{Show} as a combination of ground rewards and *weighted observer belief changes* in the true ground MDP, M_i:

$$R_i^{\text{Show}}(s_t, b_t^{\text{Obs}}, a_t, s_{t+1}, b_{t+1}^{\text{Obs}}) = R_i(s_t, a_t, s_{t+1}) + \kappa(b_{t+1}^{\text{Obs}}(M_i) - b_t^{\text{Obs}}(M_i)), \quad (3)$$

where κ controls the degree of a demonstrator's motivation to show. The showing transition function is similarly determined by the ground dynamics, T_i, as well as how an observer's beliefs change in response to observed actions:

$$T_i^{\text{Show}}(s_{t+1}, b_{t+1}^{\text{Obs}} \mid s_t, b_t^{\text{Obs}}, a_t) = \begin{cases} T_i(s_{t+1} \mid s_t, a_t), & \text{if } b_{t+1}^{\text{Obs}} = \text{BU}(s_t, a_t, s_{t+1}, b_t^{\text{Obs}}) \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

A showing demonstrator can then be modeled by calculating the solution to an OBMDP.[1] That is, for M_i^{Show}, we can calculate a value function and softmax policy that defines actions to take in different world-belief states: $\pi_i^{\text{Show}}(a_t \mid s_t, b_t^{\text{Obs}}) \propto \exp\{Q_i^{\text{Show}}(s_t, b_t^{\text{Obs}}, a_t)/\tau^{\text{Show}}\}$, where τ^{Show} is a showing temperature parameter.

Learning from Showing

Just as a showing demonstrator has a nested model of an observing agent, we can define a *sophisticated observer* who reasons about showing demonstrators. Analogous to Equation 2, a sophisticated observer updates a distribution over OBMDPs by reasoning about possible showing agents:

$$b_{t+1}^{\text{S-Obs}}(M_i) \propto \pi_i^{\text{Show}}(a_t \mid s_t, b_t^{\text{Obs}}) T_i(s_{t+1} \mid s_t, a_t) b_t^{\text{S-Obs}}(M_i). \quad (5)$$

A sophisticated observer recognizes that actions are, in part, pedagogically motivated. For instance, when teaching a child how to tie their shoes, a parent might fold the laces to clearly resemble "bunny ears". A naïve observer would only be able to attribute those particular actions to task-related goals, whereas a sophisticated observer could also reason about them in relation to communicative goals.

Summary

Here, we have developed a framework for modeling demonstrator behavior and observer interpretations of behavior based on recursive social reasoning and Theory of Mind. Different types of demonstrators and observers correspond to different "levels" in a cognitive hierarchy, as illustrated in Figure 1, allowing us to simultaneously model actions and inferences about possible tasks as they unfold over time. An implementation of the different models is available at `https://github.com/markkho/demonstration-teach-learn`.

Modeling Showing as Planning in Observer Belief Space

Task

In Ho et al. (2016), we compared how people show a task to how they do a task. Participants were given the gridworld in Figure 2a, where they could move the blue agent up, down, left, or right. Each round began in the same location and ended upon reaching the yellow goal, worth 10 points. Also, on each round, the reward for stepping on the remaining color tiles (orange, purple, and cyan) changed. Each color could be

[1] Although both plan over beliefs, an OBMDP is *not* equivalent to a Partially Observable Markov Decision Process (POMDP) with the true MDP as the hidden state. Due to Equation 4, an OBMDP value function is not necessarily piecewise, linear, and convex, which is a key property of a POMDP (Kaelbling, Littman, & Cassandra, 1998). We approximate the OBMDP value function using value iteration (Sutton & Barto, 1998) over a discretization of the belief space.

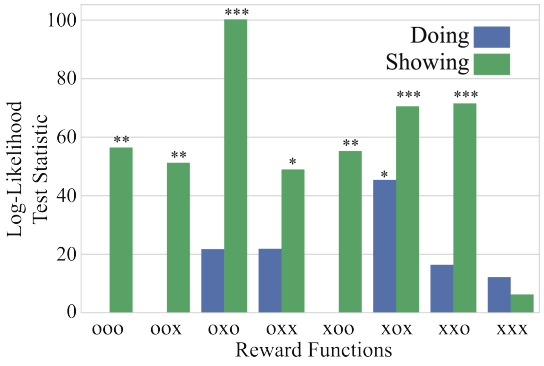

Figure 3: Tests for Ho et al. (2016) data. See Figure 2 for reward function codes. $^{*}p < .05, ^{**}p < .01, ^{***}p < .001$.

either safe (no points) or dangerous (-2 points), resulting in eight distinct reward functions (Figure 2b).

In the test phase, participants were given a round with each of the eight reward functions and were assigned to the *doing* or *showing* condition. Those in the *doing* condition were simply told the reward values of each color. Those in the *showing* condition were told the reward values, but they were also told that their behavior would be shown to another participant. Critically, this other participant would need to know which colors were safe and dangerous for a *separate* experiment. Additional experimental details can be found in Ho et al. (2016).

Analysis of Ho et al. (2016) Results

Does the current model explain showing? In an OBMDP, the value of showing comes from the rewards associated with transitions in an observer's belief space. Thus, a model of doing the task is a subset of showing the task, and we can use a likelihood-ratio test to determine if showing explains behavior. We fit the current model to individual participants and rounds, varying γ^{Do}, τ^{Do}, γ^{Show}, and τ^{Show}. Since an OBMDP collapses into the original world MDP when actions are uninformative, we used $\tau^{\text{Do}} = 1000$ as the null model in a likelihood-ratio test. Using this test on the original data, we found that for the showing condition, seven out of eight reward functions rejected the null model, while in the doing condition, only one out of eight rejected it (all $\chi^2(29) > 42.5$, $p < .05$, Figure 3).

The model of showing as planning in observer belief space thus provides an account of how peoples' showing demonstrations unfold over time. This represents several advances over previous accounts such as that presented in Ho et al. (2016). First, it directly integrates non-communicative rewards, such as losing points for being on certain tile colors, and communicative goals through R^{Show} (Equation 3). This allows us to model how people balance these motivations. Second, we can arbitrarily approximate the entire value function and policy over observer belief space, rather than be constrained to enumerated trajectories. In doing so, we can directly model extended, repetitive behaviors that are non-

Figure 4: Left: Participant demonstrations in doing condition when only orange is dangerous. Right: Showing demonstrations with an example of extended behavior captured by the model of planning in observer belief space in red.

Markov in the world state-space, but Markov in the observer belief space. For example, Figure 4 illustrates a person's showing demonstration from a single round in which specific transitions are revisited and emphasized in a particular sequence. We can also compute teaching policies when the environment involves stochastic transitions (Ho, Littman, Cushman, & Austerweil, in prep), which cannot be modeled by selecting a deterministic trajectory. Finally, by (approximately) computing a compact representation for showing policies, $\pi_i^{\text{Show}}, i = 1,2,...,n$, we can compute the belief states of a *sophisticated* observer who reasons about a demonstrator's communicative intentions ($b^{\text{S-Obs}}$). In the next section, we present an experiment designed to compare the predictions of a naïve observer and this sophisticated observer model.

Experiment: Learning from Showing

Given that the model accounts for demonstrator behavior, we can investigate how demonstrator intentions and an observer's interpretation influence what is ultimately learned. To answer this question, we presented the empirical demonstrations obtained from the doing and showing *demonstrator* conditions originally reported in Ho et al. (2016) to a new set of participants. These participants were additionally placed in either a doing or showing *observer* condition in which the interpretation of a demonstration (whether it was originally produced with the intent to show) was manipulated. Testing both sets of demonstrations as well as both possible interpretations as separate factors enables us to understand how they interact and each influence learning from demonstration.

Materials and Design

The stimuli used were the state/action/next-state tuples from the original study. These were generated from the eight main trials from the 29 participants in the doing and showing *demonstrator* conditions, for a total of 464 demonstrations. Each participant was told they would observe a single demonstration from a partner. They were also assigned to a doing or showing *observer* condition. In the showing observer condition, but not the doing one, they were told that their partner "knows that you are watching and is trying to show you which colors are safe and dangerous". Next, they were shown

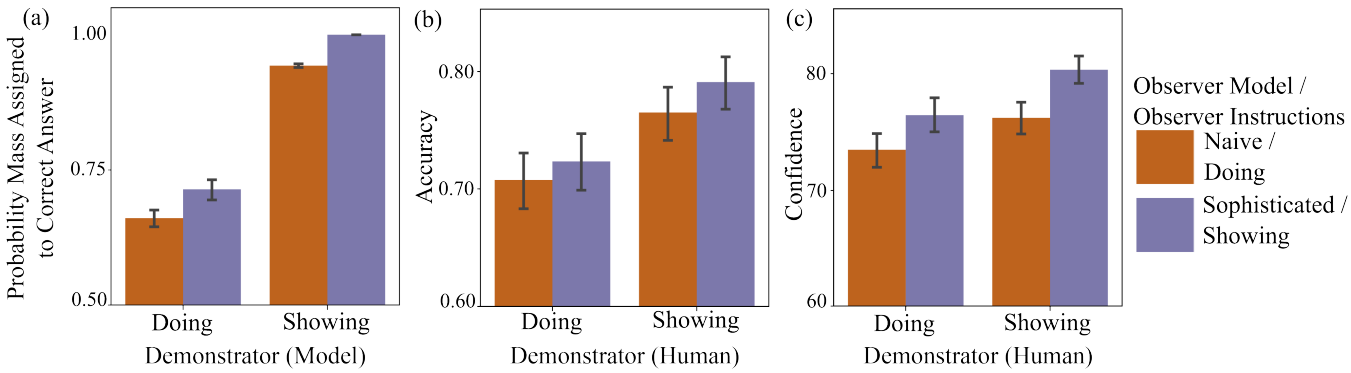

Figure 5: (a) Probability on the correct answer produced by the model. (b) Accuracy of human judgments. (c) Reported confidence of human judgments on a 0-100 scale. Error bars are bootstrap-estimated 95% confidence intervals.

a page with the animated demonstration and answered, for each of the three colors (orange, purple, and cyan), whether they thought it was safe or dangerous and their confidence on a continuous scale (0 to 100). Each participant received a base pay of 10¢ and starting from a bonus of 15¢ won/lost 5¢ for each correct/incorrect answer. Two MTurkers were assigned to each demonstration and observer instruction combination using psiTurk (Gureckis et al., 2016).

Models

We implemented four models for the gridworld task: the doing agent, the showing agent, the naïve observer, and the sophisticated observer. The doing agent was parameterized with $\gamma^{\text{Do}} = .99$ and $\tau^{\text{Do}} = .08$, while the showing model with a nested doing model was parameterized with $\gamma^{\text{Do}} = .99$, $\tau^{\text{Do}} = 3.0$, $\gamma^{\text{Show}} = .9$, and $\tau^{\text{Show}} = 1.0$. These values were chosen to produce trajectories that were qualitatively comparable to human demonstrations. For each reward function and demonstrator type, one hundred trajectories were generated, which were then fed to either a naïve observer who performed inference over possible models of doing (using the nested doing parameters) or a sophisticated observer who performed inference over possible showing models. At the end of each observation of a trajectory, we converted the final probability vector over possible MDPs to probabilities that each color was safe/dangerous. For each reward function, these were then converted to a corresponding probability on the true reward of the color that the demonstration was generated from. As shown in Figure 5a, both model showing demonstrations and sophisticated observing led to greater probability mass assigned to the correct option.

Experimental Results

For both accuracy of which colors were safe/dangerous and confidence, we found main effects of both the demonstrator and observer instructions. For judgment accuracy, we used a repeated-measures logistic regression with correct/incorrect as the outcome variable, reward function and demonstrator as random effects, and demonstrator instructions and observer instructions as fixed effects. We found significant variance across reward function intercepts ($SD = 0.92$, $\chi^2(1) = 760.26$, $p < .0001$) and demonstrator ($SD = 0.33$, $\chi^2(1) = 63.10$, $p < .0001$). The most complex model with a significant increase in fit was one with the demonstrator and observer instruction conditions as main effects, but without their interaction. In the final model, there was a main effect of demonstrator instructions ($\beta = 0.40$, $SE = 0.11$, $z = 3.63$, $p < .001$), corresponding to showing demonstrations increasing accuracy by 1.5 times, holding other factors at fixed values. There was also a main effect of observer instructions ($\beta = 0.13$, $SE = 0.07$, $z = 1.97$, $p < .05$), corresponding to observers' interpretation of demonstrations as intentional showing increasing accuracy by 1.14 times.

We similarly analyzed the confidence judgments provided by participants using a mixed-effects linear regression model. Confidence on a 0 to 100 scale was the outcome variable, while reward function, demonstrator, and observer were random effects, and demonstrator and observer instructions were fixed effects. We found significant variance across reward function intercepts ($SD = 3.65$, $\chi^2(1) = 91.97$, $p < .0001$), demonstrator ($SD = 1.20$, $\chi^2(1) = 23.72$, $p < .0001$), and observer ($SD = 14.14$, $\chi^2(1) = 454.31$, $p < .0001$). In the final model, which did not include the interaction between demonstrator and observer instructions, there was a main effect of demonstrator instructions ($\beta = 3.34$, $SE = 0.93$, $t(57.2) = 3.59$, $p < .001$) and observer instructions ($\beta = 3.57$, $SE = 0.87$, $t(1790.8) = 4.08$, $p < .001$). In short, across both measures (accuracy and confidence), we found main effects of demonstrator and observer instructions (Figure 5bc).

Discussion

We presented a computational framework for modeling demonstrator behavior and observer interpretation based on Theory of Mind (Baker et al., 2009) and recursive social reasoning (Camerer et al., 2004). In our models, the meaning of actions is grounded in what an agent performing a task would do, and a showing demonstrator is modeled as planning in the belief space of a naïve observer using Theory of

Mind. A sophisticated observer is then one who also reasons about the communicative goals of a showing demonstrator to draw stronger inferences about what they are being shown. This model has a number of advantages over one originally presented in Ho et al. (2016), and we found that it captures new aspects of the data in that study. Further, we can model the inferences of a sophisticated observer. In an experiment that used previously collected demonstrations, we found that, consistent with our models, both the observer's interpretation of behavior as showing and demonstrator's communicative intent to show positively influence learning.

Our approach draws on a number of existing ideas and relates to several other lines of research. Related formalisms have been explored in the context of making robot actions legible (Dragan, Lee, & Srinivasa, 2013) and from a "value alignment" perspective (Hadfield-Menell, Russell, Abbeel, & Dragan, 2016). Within cognitive science, this work builds on models of concept teaching by example (Shafto et al., 2014) and sequences of teacher interventions (Rafferty, Brunskill, Griffiths, & Shafto, 2016) as recursive reasoning and partially observable planning, respectively. Additionally, similar models have been used to study how people generate and interpret pragmatics in language (Frank & Goodman, 2012). This work can be seen as a direct extension of the work in pragmatics to forms of *non-verbal* communication where the "semantics" of communicative behaviors are determined by world-directed intentional action (i.e. doing tasks).

There are several directions to explore with these models. For instance, they make predictions about the time course of naïve versus sophisticated observer inferences, but our experiment did not test these directly. Important differences might arise in more complex domains with longer time horizons. Also, we model belief space transitions as deterministic and known with certainty, but in reality this is rarely the case. "Uncertainty in the observer's uncertainty" could have an influence on demonstrator behavior that cannot be explained by the current model. Finally, some work in linguistics explores the back-and-forth of conversations from the perspective of recursive social reasoning (Hawkins, Stuhlmüller, Deegan, & Goodman, 2015). This work could be extended to model situations in which both the teacher and learner can take actions while observing and reasoning about one another. Future work will need to explore these questions to provide a clearer picture of everyday teaching, social learning, and communication.

Acknowledgments

MKH was supported by a Brown University Dissertation Fellowship.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. doi: 10.1016/j.cognition.2009.07.005

Boyd, R., Richerson, P. J., & Henrich, J. (2011). Colloquium Paper: The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*(Supplement_2), 10918–10925. doi: 10.1073/pnas.1100290108

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 861–898.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. doi: 10.1016/j.tics.2009.01.005

Dennett, D. C. (1987). *The intentional stance*. MIT press.

Dragan, A., Lee, K., & Srinivasa, S. (2013). Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 301–308). doi: 10.1109/HRI.2013.6483603

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998. doi: 10.1126/science.1218633

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193. doi: 10.1016/0010-0277(95)00661-H

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative Inverse Reinforcement Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 3909–3917). Curran Associates, Inc.

Hamlin, K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental Science*, *16*(2), 209–226. doi: 10.1111/desc.12017

Hawkins, R. X., Stuhlmüller, A., Deegan, J., & Goodman, N. D. (2015). Why do you ask? Good questions provoke informative answers. In D. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 878–883). Austin, TX: Cognitive Science Society.

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243–259. doi: 10.2307/1416950

Ho, M. K., Littman, M., Cushman, F., & Austerweil, J. L. (in prep). *Teaching with and learning from demonstrations.*

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 3027–3035). Curran Associates, Inc.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1), 99–134. doi: 10.1016/S0004-3702(98)00023-X

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (1st ed.). New York, NY, USA: John Wiley & Sons, Inc.

Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2016). Faster Teaching via POMDP Planning. *Cognitive Science*, *40*(6), 1290–1332. doi: 10.1111/cogs.12290

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. doi: 10.1016/j.cogpsych.2013.12.004

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Cambridge, Massachusetts: Harvard University Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., et al. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(5), 675–690.

Wunder, M., Kaisers, M., Yaros, J. R., & Littman, M. (2011). Using iterated reasoning to predict opponent strategies. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2* (pp. 593–600). International Foundation for Autonomous Agents and Multiagent Systems.