Exploring How Information Shapes Human Inference from Demonstrations

Inga Ibs (inga.ibs@tu-darmstadt.de)
Centre for Cognitive Science, Technical University of Darmstadt

Jaron Schäfer (jaron.schaefer@stud.tu-darmstadt.de)
Centre for Cognitive Science, Technical University of Darmstadt

Constantin A. Rothkopf (constantin.rothkopf@cogsci.tu-darmstadt.de)
Centre for Cognitive Science, Technical University of Darmstadt

Abstract

Humans are remarkably proficient in discerning objectives and task nuances from observing others' behavior. This skill enables people to efficiently learn solutions for tasks from demonstrations. However, one strategy may be to simply imitate the observed actions while an alternative strategy is to infer the goals implicit in the observed behavior. In this study, we designed a navigational task to investigate computational accounts of how humans decide between the two strategies. Our primary objective is to investigate how the information content of provided demonstrations relates to the different ways humans infer an agent's policy, based on either inferring the agent's objectives or determining the agent's actions based on similar situations. Our results challenge prior findings and provide new perspectives for future research.

Keywords: Inverse Reinforcement Learning; Decision-Making; Learning from Demonstrations

Introduction

Humans are remarkably good at learning how to solve tasks by merely observing the behaviors of others. Yet, how they infer what to do in situations not shown to them might differ depending on the circumstances. Imagine learning to make tomato soup based on your family recipe. At first, you might copy each step you see. But with experience, you might start to understand why each step is important. Following the recipe exactly is an easy way to make a good tomato soup, but knowing why things are done helps to adapt the recipe in new contexts, e.g., if ingredients differ slightly. Studies in social cognition about theory of mind suggest that humans build mental models of others based on their behavior by attributing goals and utilities to their actions (Baker, Saxe, & Tenenbaum, 2009; Jara-Ettinger, 2019; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016), modeled by inverse reinforcement learning (IRL). However, recent observations indicate that in some situations, humans compare actions across similar states when reconstructing a policy, as described by imitation learning (IL) models (Lage, Lifschitz, Doshi-Velez, & Amir, 2019). Which type of inference is preferred in which situation is not clear. Here, we hypothesize that the choice of inference depends on the complexity of inference of the reward function, which captures the relationship between environment features and objectives. In tasks where the reward function consists of many complex features that are not easily discernible from information in the demonstrations, it might be preferable to imitate the behavior as closely as possible. In cases where the demonstrations do not provide enough similar states to conclude the action, it might be better to infer the actual goals from the demonstrations rather than to imitate directly. Here, we propose an experimental paradigm for distinguishing different types of inference from demonstrations. We investigate whether humans deploy different inference types dependent on measures of information content in demonstrations for each type of inference.

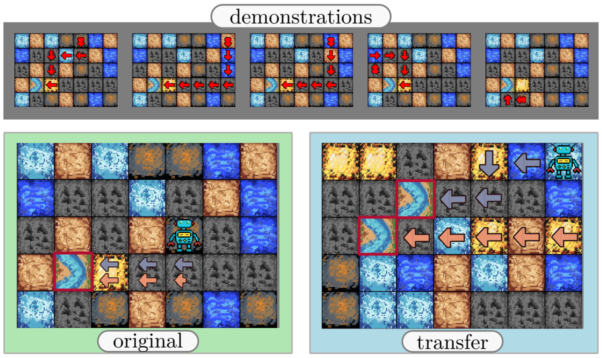

Figure 1: Task environment. Each trial showed five demonstrations at the top, first the original and then the transfer world. Arrows in blue and orange show the predictions of the IRL and IL models, respectively, and the red squares indicate the target areas (not shown to participants).

Methods

Experimental Design Participants (N=17) were asked to navigate a robot on a foreign planet, represented by a 7×5 grid world as is shown in Figure 1, to collect a sample from a specific area. Each grid cell was associated with different fuel costs or the target area, distinguishable by their visual appearance. The costs and the target area were unknown, but actions of an optimal policy were shown in five demonstrations, illustrating routes from five distinct starting points to one of the target cells. The participant's task was to move the robot to a target area on the most cost-efficient route based on the information shown in the demonstrations. The navigation task can be formalized by a Markov Decision Process (MDP) in a tuple (S, A, T, R), with S as the finite set of states (cells in the grid world), ϕ_s the feature vector of a state s (appearance of the cell), and A the finite set of actions (four-movement directions). The transition function T is deterministic, such that the agent moves to the neighboring field in the direction encoded by the action. R is the reward function that associates each state with a reward, dependent on its appearance. Regular (gray) states give a reward of -1; target cells a reward of 100; and other states reward between -1 and -20. To vary the complexity of the reward function, we varied the number of features, the density of non-regular states, and the number of features associated with the reward of -1. In each of the 32 trials, participants had to navigate the robot in two conditions. One in which the layout of the features matched the layout shown in the demonstration (*original condition*). In a second condition, the *transfer condition*, participants were shown the same demonstrations; however, the world they had to navigate the robot differed from the original layout. Importantly, the rewards associated with the appearance of the cells stayed the same. The transfer condition allowed us to gather participants' decisions from initial states where different inference models predicted distinct actions (see Figure 1).

Models of Inference To ensure comparability with the study of Lage et al. (2019), we use the same models for IRL-based and IL-based inference. As a model for inverse reinforcement learning, we use Maximum-Entropy Inverse Reinforcement Learning (Ziebart, Maas, Bagnell, Dey, et al., 2008), which infers a reward function by matching the expectation of reward features of a candidate policy to the ones observed in trajectories of an agent. We employ a Gaussian Random Field (GRF) classifier as an IL-model, which predicts actions in states based on comparable states as proposed in Zhu, Lafferty, and Ghahramani (2003). The features provided to the classifier correspond to the feature of the state the agent is in, as well as the features of the four surrounding cells. As Lage et al. (2019), we use the extension of the GRF to the multi-class setting, where a one-vs-rest classification is performed for each class. For model comparison, we map the classification results to values between zero and one using a softmax function on the class margins.

Choosing Demonstrations In each trial, we chose five demonstrations, which either maximized the information for inference with the IRL model or the inference with the IL model using two algorithms also proposed by Lage et al. (2019). To maximize the information for the IRL model, we used the SCOT machine teaching algorithm (Brown & Niekum, 2019), which selects the demonstrations that constrain all reward functions to functions with associated behaviorally equivalent policies to the optimal policy, its behavioral equivalence class (BEC). The SCOT algorithm greedily chooses the demonstrations that cover most of the constraints of the policy's BEC class. To maximize the information for IL-based inference, we used an active learning approach proposed originally by Zhu et al. (2003) with the adaptation used in Lage et al. (2019), choosing demonstrations that minimize the prediction loss of the GRF on states not included in demonstrations.

Information Content We define a measure for the information content the demonstrations provide for each model. The information content for the IL model represents how well states in the demonstrations allow the comparison with other states. We define it as the prediction loss of the IL model on the unlabeled states in the original condition. For the IRL-based inference, the information content is described by the proportion of constraints covered of the BEC of a policy in the original condition by demonstrations. It relates to the uncertainty over possible reward weights the IRL algorithm could infer.

Results

Figure 2 a) shows the proportion of suboptimal moves made by each model and participants in each condition. The IRL model and the participants outperform the IL model in both conditions. The IL model performs worse in the original than in the transfer condition, whereas both the IRL and the human performance remain stable over both conditions. Model comparison of the human trajectories with each model (see Figure2 b)) shows that in both conditions, the IRL model generally matches the participants' decisions better than the IL model. Figure 2 c) shows the likelihood ratio of the IL and the IRL model for each trajectory in the transfer condition. The trajectories are grouped by the information content in the demonstrations. The IRL model better models most trajectories; however, some trajectories, 19 of the 224 trajectories, are better explained by the imitation learning model, 16 of them in trials with a high number of features (more than seven).

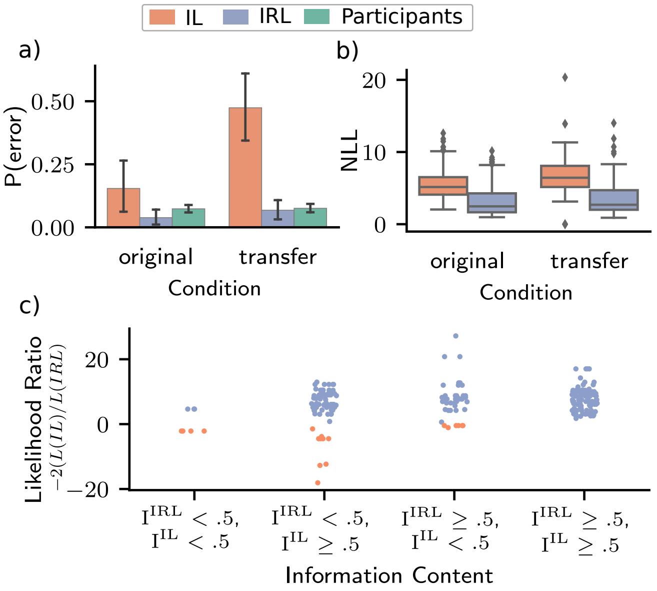

Figure 2: Comparison of model and human behavior. a) Proportion of suboptimal moves of the two inference models and the participants for each condition. b) Negative log-likelihood scores for the participants' trajectories under each model. c) Likelihood ratio of the models on individual trajectories in the transfer condition, grouped by information content. I^{model} represents the information content for each model.

Discussion & Conclusion

We introduced an experimental paradigm to investigate circumstances in which humans prefer different types of inference. In a first study, we investigated whether participants switch between IL-based and IRL-based inference based on the information content of the demonstrations. Unlike the observations of Lage et al. (2019) in experiments utilizing similar tasks, we did not see a prominent use of IL-based inference. One explanation could be that participants did not gain any advantage from IL-based inference in the transfer task since it performs worse on data that differs from the original distribution. However, the results suggest that it might be preferred in cases of low information content for at least one of the inference types and a high number of features. In the future, we plan to investigate models of Bayesian inference and other IL-based inference models by strategically designing trials to cover the information values for the respective models equally.

Acknowledgments

We thank Dominik Straub and Vildan Salikutluk for the helpful discussions and comments on this manuscript. This work was funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) in the project 'PlexPlain: Erklärende KI für komplexe Lineare Programme am Beispiel intelligenter Energiesysteme' [grant number 01IS19081] and by the Hessian Ministry of Higher Education, Research, Science and the Arts and its LOEWE research priority program 'WhiteBox' under grant LOEWE/2/13/519/03/06.001(0010)/77.

References

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

Brown, D. S., & Niekum, S. (2019). Machine teaching for inverse reinforcement learning: Algorithms and applications. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 7749–7758).

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

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Lage, I., Lifschitz, D., Doshi-Velez, F., & Amir, O. (2019). Exploring computational user models for agent policy summarization. In *Ijcai: proceedings of the conference* (Vol. 28, p. 1401).

Zhu, X., Lafferty, J., & Ghahramani, Z. (2003). Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *Icml 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining* (Vol. 3).

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. (2008). Maximum entropy inverse reinforcement learning. In *Aaai* (Vol. 8, pp. 1433–1438).