Towards Online Adaptation for Autonomous Household Assistants

Benjamin A. Newman
newmanba@cmu.edu
Carnegie Mellon University, Meta AI
Pittsburgh, Pennsylvania, USA

Christopher Jason Paxton
cpaxton@meta.com
Meta AI
Pittsburgh, Pennsylvania, USA

Kris Kitani
kkitani@cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Henny Admoni
henny@cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

ABSTRACT

Many assistive home robotics applications assume open-loop interactions: robots incorporate little feedback from people while autonomously completing tasks. This places undue burden on people to condition their actions and environment to maximize the likelihood of their desired outcomes. We formalize assistive household rearrangement as collaborative online inverse reinforcement learning (IRL). Since online IRL can lead to sample inefficient interactions and overfit to specific user objectives, we compare sample efficiency and generalizability of two initial choices of action representations in a simulated household rearrangement task. We show, under certain assumptions, that representing objects by their material properties can increase sample efficiency and generalizability to out of domain objects.

CCS CONCEPTS

• **Computing methodologies** → **Cooperation and coordination**; **Learning from implicit feedback**; **Inverse reinforcement learning**.

KEYWORDS

assistive robotics, object rearrangement, online inverse reinforcement learning, household robots

ACM Reference Format:
Benjamin A. Newman, Christopher Jason Paxton, Kris Kitani, and Henny Admoni. 2023. Towards Online Adaptation for Autonomous Household Assistants. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580136

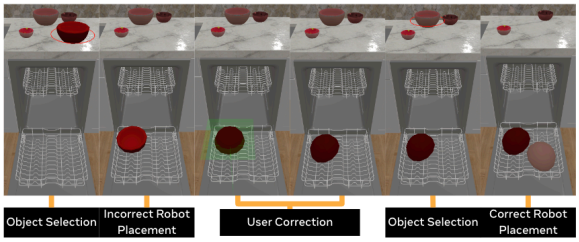

Figure 1: One step of a surface rearrangement problem: dishwasher loading. From left to right: a person picks a bowl to place in the dishwasher, the robot places this incorrectly, this is then corrected by the person placing the bowl concave side down. The robot learns that the person likes bowls placed with the concave side down and is able to place the next bowl correctly.

1 INTRODUCTION

Assistive home robots predominately operate in open-loop paradigms: robots autonomously complete household chores after being issued explicit commands by an operator (or according to a predetermined routine) without incorporating feedback from the person during task completion. Prior approaches rely on preprogrammed routines or operator teleoperation [8] to find solutions to long-horizon household tasks that have complex temporal dependencies and ambiguous optimality. Recent research resolves this using deep neural networks, such as large language models [2]. Both approaches, though, neglect to learn from human feedback about robot performance.

While these interaction designs may work well in laboratory settings, they are unlikely to extend to *in situ* interactions with novice robot operators whose preferences vary from preprogrammed subroutines or the mode of a training data distribution. To express their preferences, open-loop collaborations require people to explicitly choose actions or states that include the necessary context for their preferences to be carried out. This can result in cumbersome behaviors that are difficult for people to exhibit, such as overly descriptive natural language instructions or full task demonstrations. These actions may also be impossible to produce when a preferences remain latent until discovered through exposure to a specific stimulus. This interaction design not only introduces additional burden on people, but it neglects valuable sources of information already being disclosed through people's goal oriented behavior. Due to this increased burden of action production, relying on action and state conditioning to execute assistive robot

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03.
https://doi.org/10.1145/3568294.3580136

actions may lead to robot policies that are inconsistent with the presupposed assistive role of the robot [13].

To address this, we suggest that robots interacting with people in their homes should incorporate feedback from naturally exhibited, collaborative user behavior to ensure assistive robotic behaviors adhere to people's preferences. To test this, we formulate household collaborations as rearrangement problems [5, 17] solved through online IRL making use of people's naturalistic behavior. This setup has been shown to work well when people's behavior is interpreted as feedback about low-level robot actions, such as in shared control [10] and autonomous robot path planning [12], by employing maximum entropy inverse reinforcement learning (MaxEntIRL) [4, 19]. Instead of interpreting people's goal-directed behavior as feedback about a robot's low-level actions, we assume low-level robot policies and interpret people's behavior as information about the high-level plan the robot should execute.

One potential drawback of this method is sample inefficiency. Depending on the frequency with which people act and the correlation between these actions and people's goals, it can require many episodes before converging to a desired solution. We show how choosing appropriate action and task objective representation spaces can increase the sample efficiency by, under certain assumptions about people's task objectives, converging more quickly to a ground truth objective and generalizing to actions not previously observed during training.

Our goal with this work is to present a first step towards developing assistive household robots that incorporate user feedback into their high-level task plans in a manner consistent with the nature of assistive relationships. We first present an extension of online IRL techniques typically used in short-horizon, low-level robot control tasks to long-horizon, sequentially dependent, high-level household rearrangement tasks. Then we explore how the choice of action and objective representations can make more efficient use of a person's goal oriented behaviors during task execution.

2 RELATED WORK

We first present work in state and action-conditioned models that do not explicitly learn about their human partner. Then, we review work in online adaptation.

2.1 State and action-conditioned collaboration

Zero-shot coordination is a recent field of research aiming to develop collaborative agents that can successfully and immediately interact with new people. This is typically done by pretraining models in simulation against agents designed to mimic human behavior [6] or over a diverse population of simulated agents [16]. Large language models that propose task plans executed by robots [2] have also been used for assistive collaboration in household tasks. These methods do not continually adapt to individual users, who may not fit well within the distribution seen during training. In this work we focus on explicitly adapting the robot's policy to an individual, instead of placing the burden on the person to alter their behavior to achieve a more favorable robot policy. Sophisticated offline pretraining techniques, such as those used in zero-shot coordination, could be used to initialize adaptive models in future work.

2.2 IRL for adaptive collaboration

Using IRL for robot control can be difficult, in part, due to the ambiguity that arises from traditional IRL [1]. Maximum entropy IRL facilitates this by using the principle of maximum entropy to order solutions according to how well they match observed user behavior [19]. This solution has also been used in behavioral science to model people's ability to infer others' goals from their behavior exhibited during goal-directed plans [4].

These insights have been applied to robot trajectory optimization for shared control. In the difficult task of teleoperating a high-degree of freedom robot arm with a low-degree of freedom input device, such as a joystick, a robot can assist by observing user input commands, inferring the most likely goal from a set of predetermined goals, and moving along a path towards this goal [10]. MaxEntIRL can also been used to interpret less direct forms of user behavior, such as a person physically pushing a robot to express their preference for the robot's trajectory when, for example, carrying a coffee mug around a laptop computer [12]. We adapt online MaxEntIRL for determining high-level task plans that are consistent with user behavior in household collaborations.

IRL has also been applied to learn robot policies in other types of human-robot interactions. For example, to learn people's preferences from observations of independent task demonstrations [18], or by learning assistive social actions for therapy combining a therapists' expertise with their demonstrations [3], or for social health, as a robot receptionist learning to give hygiene advice in a shopping mall [7]. Our formulation learns preferences from in-situ, collaborative behavior, for collaborative rearrangement tasks, ideally minimizing the distribution between interaction data seen at training and test time.

3 METHODS

We are interested in developing robots that assist people while completing collaborative household tasks such as dishwasher loading or table setting. First, we formalize the task of surface rearrangement, a specific instance of rearrangement problems [5, 17], as a decentralized partially observable Markov decision problem (DEC-POMDP). Then, we present an algorithm for solving such tasks.

3.1 Problem Setup

To study online adaptation for assistive agents in tightly knit collaborations, we model a task which we call *surface rearrangement*. Unlike general rearrangement problems, we assume objects can be instantaneously grabbed and placed, removing the need for navigation and allowing us to focus on the collaboration.

Given this description, we can model surface rearrangement as a DEC-POMDP which is a tuple of $(S, A, U, T, Z, O, r, \gamma)$ where:

- S is the set of possible states. As in prior work, we assume that the state is a tuple of observable and unobservable features of the task $(x, \{\theta_i\})$. In our formulation, x describes the observable environment and θ_i describes the objective for each agent $a_i \in A$, where θ_i is not observable by an agent a_j when $i \neq j$.
- A is the set of agents. We assume two agents, initially.
- U_i is the set of actions for a particular agent a_i.
- Z_i is the set of observations for agent a_i.

- $T(s^t, \mathbf{u}, s^{t+1})$ denotes the transition dynamics. As in prior work, these dynamics are dictated by θ, which we assume to be constant over time.
- $O_i(s^{t+1}, u_i^t, z^{t+1})$, the observation distribution for agent a_i.
- $r_i(s^t, \{u_i\}^t)$ is the reward function for the system. In assistive settings, we assume this is equivalent to the person's reward function.
- γ is the discounting factor.

We assume two agents, one of which is a person over whose policy we have no control. Given this, we can reduce the DEC-POMDP to a single agent POMDP. Prior work in online human robot collaboration [12] shows how this POMDP can be solved using the QMDP approximation, [10, 11] and then using online gradient descent [12]. We adapt this for collaborative assistance in Alg. 1.

3.2 Surface Rearrangement

Algorithm 1 Online Learning for Assistive Surface Rearrangement

Require: $\theta, \hat{\theta}^0, \phi_{\text{obj}}, \phi_{\text{loc}}, \pi_l, \pi_f$, env,$\alpha, \gamma$

$s^0 \leftarrow$ env.reset()
placed_objects $\leftarrow$ []
while placed_objects.len() < env.objects.len() **do**
  $u^t_{\text{obj}}, \hat{u}^t_{\text{loc}} \leftarrow \pi_l(\cdot|s^t; \theta)$
  $\hat{r}^t_l \leftarrow \phi_{\text{obj}}\left(u^t_{\text{obj}}\right) \cdot \theta \cdot \phi_{\text{loc}}\left(\hat{u}^t_{\text{loc}}\right)$
  $u^t_{\text{loc}} \leftarrow \pi_f\left(\cdot|s, u_{\text{obj}}; \hat{\theta}^t\right)$
  $s^{t+1} \leftarrow$ env.step $\left(u^t_{\text{obj}}, u^t_{\text{loc}}\right)$
  $r^t_l \leftarrow \phi_{\text{obj}}\left(u^t_{\text{obj}}\right) \cdot \theta \cdot \phi_{\text{loc}}\left(u^t_{\text{loc}}\right)$
  if $\hat{r}^t_l > r^t_l$ **then**
    $s^{t+1} \leftarrow$ env.step $\left(u^t_{\text{obj}}, \hat{u}^t_{\text{loc}}\right)$
  else
    $\hat{u}^t_{\text{loc}} \leftarrow u^t_{\text{loc}}$
  end if
  $\nabla\Phi \leftarrow \phi_{\text{obj}}\left(u_{\text{obj}}\right) \cdot \phi_{\text{loc}}\left(\hat{u}^t_{\text{loc}}\right) - \phi_{\text{obj}}\left(u_{\text{obj}}\right) \cdot \phi_{\text{loc}}\left(u^t_{\text{loc}}\right)$
  $\theta^{t+1} \leftarrow \theta^t + \alpha \cdot \gamma^t \cdot \nabla\Phi$
  placed_objects.append $\left(u_{\text{obj}}\right)$
end while

We aim to model close collaborations among multiple agents performing household tasks. We model two agents partaking in a modified, collaborative pick and place task. We call one agent the "leader" and the other the "follower". The behavior for each is dictated by a reward function:

$$r_i(x, u_i, u_j; \theta_i) = \phi_i(x, u_i) \cdot \theta_i \cdot \phi_j(x, u_j).$$

The goal of both agents is to arrange the objects onto the surface in a manner that maximizes the leader's reward function, r_l, which is initially unknown to the follower. The follower must infer the parameters of r_l, θ_l, through the behavior of a_{leader}.

The leader a_l selects an object, u_{obj}, and the a_f selects a location, u_{loc}. In practice either agent could play either role. The leader selects u_{obj} such that, if placed correctly, it would maximize r_l, thereby calculating an implicit reward known only to a_l. The object is passed to a_f who selects u_{loc} to maximizes r_f, given the current state and estimate of the leader's objective: $\theta^t_f = \hat{\theta}^t_l$. Once u_{loc} is executed, a_l can correct this action if the reward received for executing u_{loc} is less than the implicit reward calculated when selecting u_{obj}. The follower sees the correction and uses the differences in the features between the two states (i.e. where a_f initially place the object and where a_l placed it after the correction) to update its estimate of the leader's objective.

4 INITIAL RESULTS

Online IRL relies on people exhibiting behaviors that are highly correlated with their goals, often leading to action representations that do not benefit from shared structure, leading to poor sample efficiency and out of domain generalization. To study this, we consider two different choices of action representation: per ID, ϕ^{ID}, which represents actions as one-hot vectors, and per Quality, ϕ^{Quality}, which represents actions by shared properties between different actions, such as an object's material, shape or size in the dishwasher loading domain. We aim to show how ϕ^{Quality} can improve sample efficiency and generalizability when people's task objectives are correlated along their actions, for example people place glass bowls into the dishwasher in a similar manner to glass cups.

We first develop a set of simulated user objectives, θ. Eight highly correlated objectives, shown in the top row of Fig. 2 are blended with randomly sampled objectives according to $\theta = \theta_{\text{correlated}} * (1 - \beta) + \theta_{\text{rand}} * \beta$ along five correlation thresholds $\beta \in \{0, 0.01, 0.1, 0.5, 1\}$, shown in the bottom row of Fig. 2, to develop objectives with varying degrees of correlation along actions. We then train randomly initialized robot objective,s $\hat{\theta}$, to approximate these user objectives, θ, using Algorithm 1.

Figure 2: Highly correlated objectives across qualities (top) and those same objectives blended with noise ($\beta = 0.5$). Rows of an objective represent surface locations and columns represent objects. Elements represent the value associated with placing that object in that location, ranging from $[-1, 1]$.

Furthermore, we create a set of O objects that share features with another object along at least one quality, (e.g. a glass mug, a glass bowl, a plastic mug, and a plastic bowl) and represent our surface by a grid of L locations. Objects are split into training and testing sets.

Per ID representations return one-hot vectors with the element representing the object or location ID activated. For objects, this results in an action space U_{obj}, that is a square identity matrix of size $|O| \times |O|$. For locations, U_{loc} this is a square identity matrix of size $|L| \times |L|$.

Per quality representations share features across different object categories by concatenating one-hot vectors over each quality (e.g. material or shape), $q \in Q$. Per quality representations of objects returns an action space U_{obj} that is of size $|O| \times \sum_{q \in Q_{\text{obj}}} |q|$, or similarly for locations, U_{loc} of size $|L| \times \sum_{q \in Q_{\text{loc}}} |q|$. Per ID representations imply a θ of size $|O| \times |L|$, while per quality representations imply a θ of size $\sum_{q \in Q_{\text{obj}}} |q| \times \sum_{q \in Q_{\text{loc}}} |q|$. When the total size of the qualities is less than the number of objects they describe, per quality representations yield a more efficient representation of the space.

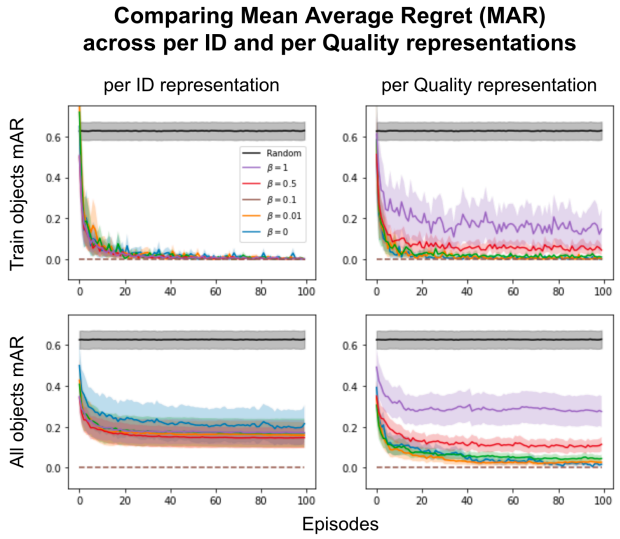

Figure 3: We report mean average regret (mAR). From left to right we show per ID representations and per Quality representations. From top to bottom, we show training and zero-shot testing on out of domain objects after N iterations. Colors indicate β thresholds as follows: 0 is blue, 0.01 is orange, 0.1 is green, 0.5 is red, and 1 is purple. Both methods perform similarly in training for well-correlated objectives, while per ID representations outperform when preferences are uncorrelated. For per Quality representations, improvements when training on correlated preferences correspond to improvements in testing set, which does not hold true for per ID representations.

Each line in Fig. 3 shows a regret curve for a separate objective, (ranging from highly correlated across qualities to randomly correlated across features). Shaded regions showing the standard error over the collection of runs. Mean average regret (mAR) for a policy taking random actions is shown in black and a reference zero-line is shown in dotted brown. Per ID results are shown on the left, per quality representations are shown on the right. Training mAR curves are shown on top, and zero-shot performance on the all objects on the bottom. Each increment along the x-axis is an episode ranging between one and six object placements.

We report regret over other potential metrics, such as accuracy or corrections, because it reflects the underlying reward our algorithm receives. Accuracy and corrections are both overly critical metrics: accuracy penalizes an algorithm for choosing an "incorrect" placement even if it returns the same reward as the "correct" label, while corrections penalize all incorrect placements equally, even when differences in reward received may be negligible.

These results show both representations capture objectives with strong correlations across object qualities and episodes of in-domain objects. For highly correlated objectives, per quality representations converge slightly faster than per ID representations, though they under perform as objectives become less correlated.

These results also show that per quality representations can generalize to out of domain objects, when objectives are highly correlated. This is because these representations can express preferences such as "place all glass objects on the top", something per ID representations are unable to do. Taken with the results on in-domain training, we show how choosing an appropriate representation space can improve upon the drawbacks of online IRL, namely that it can be sample inefficient and overfit to in-domain data.

5 LIMITATIONS AND FUTURE WORK

First, we present results validated only through simulation in low-dimensional surface rearrangement problems. We do not know how well correlated or uncorrelated people's objectives are, along which features they may be correlated, or how they interact with high-dimensional rearrangement tasks. While we are hopeful that people will have preferences that are highly correlated along object qualities, our approach needs to be validated with user studies.

Not only would human subjects studies provide an opportunity to verify our models, it would offer opportunity for improvement, as well. Collecting interaction data through interactive simulators, such as AI Habitat [15, 17], deployed as experiments on crowdsourcing platforms such as Amazon Mechanical Turk [9] or Prolific [14] would allow us to pretrain data-driven models with data about real human preferences.

Finally, we assume no cost corrections. Assigning costs to corrections would disincentivize providing low-reward corrections until the reward associated with executing corrections outweighs the execution cost, or suppress low-reward corrections altogether. This emphasizes the importance of having representation spaces that efficiently interpret information about people's objectives. We will explore learning these representation spaces from human data in future work.

Finally, our approach, also assumes people will provide direct state corrections. This maximizes the correlation between the leader's corrections and objectives. We would like to extend our approach to account for other types of corrections, such as those expressed through verbal or nonverbal communication.

6 CONCLUSION

We presented a formalization of assistive household collaborations as online IRL. This formalization is consistent with the assistive role of the robot and allows people to express preferences for tasks outcomes through naturalistic behavior. We show initial results in improving sample efficiency and generalizability of recovered rewards by choosing efficient representations. Finally, we discuss plans to extend this work to more naturalistic interactions.

REFERENCES

[1] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*. 1.

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691* (2022).

[3] Antonio Andriella, Carme Torras, Carla Abdelnour, and Guillem Alenyà. 2022. Introducing CARESSER: A framework for in situ learning robot social assistance from expert knowledge and demonstrations. *User Modeling and User-Adapted Interaction* (03 2022). https://doi.org/10.1007/s11257-021-09316-5

[4] Chris L Baker, Joshua B Tenenbaum, and Rebecca R Saxe. 2007. Goal inference as inverse planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 29.

[5] Dhruv Batra, Angel X Chang, Sonia Chernova, Andrew J Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, et al. 2020. Rearrangement: A challenge for embodied ai. *arXiv preprint arXiv:2011.01975* (2020).

[6] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems* 32 (2019).

[7] Zhichao Chen, Yutaka Nakamura, and Hiroshi Ishiguro. 2022. Android as a Receptionist in a Shopping Mall Using Inverse Reinforcement Learning. *IEEE Robotics and Automation Letters* 7, 3 (2022), 7091–7098. https://doi.org/10.1109/LRA.2022.3180042

[8] Matei Ciocarlie, Kaijen Hsiao, Adam Leeper, and David Gossow. 2012. Mobile manipulation through an assistive home robot. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 5313–5320. https://doi.org/10.1109/IROS.2012.6385907

[9] Kevin Crowston. 2012. Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars, booktitle=Shaping the Future of ICT Research. Methods and Approaches, Anol Bhattacherjee and Brian Fitzgerald (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 210–221.

[10] Shervin Javdani, Henny Admoni, Stefania Pellegrinelli, Siddhartha S. Srinivasa, and J. Andrew Bagnell. 2018. Shared autonomy via hindsight optimization for teleoperation and teaming. *The International Journal of Robotics Research* 37, 7 (2018), 717–742. https://doi.org/10.1177/0278364918776060 arXiv:https://doi.org/10.1177/0278364918776060

[11] Michael L Littman, Anthony R Cassandra, and Leslie Pack Kaelbling. 1995. Learning policies for partially observable environments: Scaling up. In *Machine Learning Proceedings 1995*. Elsevier, 362–370.

[12] Dylan P Losey, Andrea Bajcsy, Marcia K O'Malley, and Anca D Dragan. 2022. Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research* 41, 1 (2022), 20–44.

[13] Benjamin A. Newman, Reuben M. Aronson, Kris Kitani, and Henny Admoni. 2022. Helping People Through Space and Time: Assistance as a Perspective on Human-Robot Interaction. *Frontiers in Robotics and AI* 8 (2022). https://doi.org/10.3389/frobt.2021.720319

[14] Prolific. 2014 [Online]. Prolific. https://www.prolific.co

[15] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9339–9347.

[16] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with humans without human data. *Advances in Neural Information Processing Systems* 34 (2021), 14502–14515.

[17] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John M Turner, Noah D Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimír Vondruš, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel X Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=DPHsCQ8OpA

[18] Bryce Woodworth, Francesco Ferrari, Teofilo E. Zosa, and Laurel D. Riek. 2018. Preference Learning in Assistive Robotics: Observational Repeated Inverse Reinforcement Learning. In *Proceedings of the 3rd Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research, Vol. 85)*, Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (Eds.). PMLR, 420–439. https://proceedings.mlr.press/v85/woodworth18a.html

[19] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. 2008. Maximum Entropy Inverse Reinforcement Learning. In *Proc. AAAI*. 1433–1438.