# Towards Learning Interpretable Features from Interventions

Erin Hedlund-Botti
erin.botti@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Nina Moorman
ninamoorman@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Julianna Schalkwyk
jschalkwyk3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Chuxuan Yang
soyang@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Lakshmi Seelam
lseelam3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Sanne van Waveren
sanne@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Russell Perkins
russell_perkins@student.uml.edu
University of Massachusetts, Lowell
Lowell, Massachussets, USA

Paul Robinette
paul_robinette@uml.edu
University of Massachusetts, Lowell
Lowell, Massachussets, USA

Matthew Gombolay
matthew.gombolay@cc.gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

## ABSTRACT

Assistive in-home robots are one solution to help an aging population. To provide effective care, robots must be adaptable and personalizable, as everyone has different needs and preferences. We want to enable people to communicate their preferences to a robot intuitively. In this work, we propose a method where users critique the robot by intervening when the robot makes a mistake or does not follow the user's preference, in a learning from demonstration setting. The robot then learns interpretable features about the users' goals and preferences based on the intervention. We propose a series of user studies to inform and validate our framework.

## CCS CONCEPTS

• **Human-centered computing** → **User studies**; *User interface programming*; **Collaborative interaction**.

## KEYWORDS

feature learning, learning from demonstration, personalization

**ACM Reference Format:**
Erin Hedlund-Botti, Nina Moorman, Julianna Schalkwyk, Chuxuan Yang, Lakshmi Seelam, Sanne van Waveren, Russell Perkins, Paul Robinette, and Matthew Gombolay. 2024. Towards Learning Interpretable Features from Interventions. In *Proceedings of HRI (HRI '24 LEAP Workshop).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Due to an aging population, there is a shortage of caregivers [24, 38]. One solution for this shortage is to help people age in place; robots can offer assistance to enable older adults to live independently for longer. As assistive robots will be deployed in home environments, these robots must be personalizable in order to meet diverse sets of possible needs and preferences. Additionally, people's preferences will change as they age and their environments change. When a robot performs a task incorrectly or not aligned with a user's current preference, users will need to be able to quickly and intuitively correct the robot's behavior.

One method that enables non-expert users to communicate their preferences to a robot is Learning from Demonstration (LfD) [30]. In LfD, the robot learns from a recording of the human demonstrating the task. The simplest form of LfD, behavioral cloning (BC), mimics the human, learning *how* to complete the task, but not *why*. Unfortunately, BC is susceptible to covariate shift [31] and cannot adapt when the transition function changes. Inverse reinforcement learning (IRL) techniques attempt to understand context by learning a reward function for the human's goal [25]. However, IRL methods are not sample-efficient for the end-user [2, 16], and most LfD techniques are not interpretable [33]. In this work, we propose a new LfD framework that is interpretable, sample-efficient, and enables the robot to predict which features a user wants the robot to consider when learning a new behavior.

In an LfD setup, it can be difficult for people to identify and communicate the features that are important to accomplishing the task [14]. A more effective approach may be for people to critique after observing a robot's attempt to perform a desired task. We posit that when people intervene during a robot failure or provide corrective feedback, there is semantic information from which the robot can learn. For example, if someone stops the robot when the robot is close to colliding with an obstacle, the robot could learn that the important feature is to keep a "safe" distance from obstacles. Furthermore, when the person interrupts (e.g., how close the robot is to the obstacle) can provide the robot with what a "safe" distance means for that person. We propose to have the robot attempt the task and have participants intervene when the robot behavior does not match their preference. From these interventions, we will learn

This work was supported by NSF IIS-2112633 and the PEO Scholar Award.


*HRI '24 LEAP Workshop, March 11, 2024, Boulder, CO*

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

interpretable features that can help the robot understand environmental and temporal context clues to improve the robot's policy to better account for human preferences.

Our goal is to provide users with an efficient and intuitive method for personalizing robot behavior that can help robots adapt as people's lives change over their lifetimes. In this paper, we propose a multi-study design to learn and evaluate a model for learning interpretable features from interventions. First, we aim to conduct a pilot study to gather a list of features from participants and utilize this information to develop a user interface. Then we will conduct a data collection study to train the framework. Participants will watch the robot attempt a series of household tasks and intervene when the robot does not behave as expected or desired. After intervening, participants will label the feature of importance using the interface. Then participants will provide a demonstration of the correct behavior to the robot. After training the model, we aim to evaluate our framework and features compared to baselines (i.e. without features, behavioral cloning, and ablations of our method). We plan to learn generalizable features that can be used to correct and improve a robot's behavior to best suit the end user.

## 2 RELATED WORKS

LfD seeks to enable humans to teach robots new skills via human task demonstrations without the need for users to have programming experience [30]. Robot-centric LfD learns from a human's corrective feedback as the robot executes the task in the form of action corrections [32] or scalar feedback [15, 21]. Our work is inspired by Kelly et al. [20] and Spencer et al. [37] who proposed having people take over task execution when the robot deviates from the desired behavior. We go beyond this prior work, learning to predict semantically meaningful features from the interventions.

Our approach for LfD addresses the reality that robots will not know a priori which features a user cares about when specifying a new behavior via demonstration. With a feature mismatch, the robot will fail to learn the skill, which can result in trust degradation [17, 19, 34]. In our work, we focus on system and design errors [39] to inform how we should design our feature-specification interface for LfD. Additionally, we investigate ambiguous or subjective failures, which are based on people's preferences [26].

LfD researchers have considered various approaches for learning from robot failures. One solution is to request demonstrations from the user to learn how to recover [8] or allow the user to intervene [10, 20, 28, 37]. Other methods learn constraints from the demonstrations to learn "safe" boundaries for the policy [23, 27]. Our work is complementary to prior work as we aim to use the information learned from the robot's failure and the human's intervention to determine features of importance and improve the robot's policy.

Prior work has also investigated learning what feature prompted an instance of user feedback [6]. In such a setting, Bajcsy et al. [4] show that learning one feature per intervention compared to all at once results in improvements in objective and subjective results. Learning-based approaches that use neural networks are not guaranteed to be interpretable to a non-expert user; however, interpretability offers the benefit of transparency and agent accountability [33]. We endeavor for our features to be readable, human-worded, and understandable [40]. Das et al. [11] expand on learning interpretable features, showing that presenting both the context of the failure and preceding robots actions to be helpful. As prior work has found that querying people affords interpretable, relevant features and that contrasting examples are key to feature specification, we obtain our features directly from participants by soliciting feedback as they observe the robot attempt the task [9].


**Figure 1: This figure depicts our four domains.**

## 3 METHODS

In this work, we design a framework that learns users' preferences for how a robot should complete a task based on user intervention during failure. An overview of our experiment procedures is in Figure 2. We will first conduct a pilot study to evaluate our experimental design and elicit a list of relevant features from users. Then we will conduct a data collection study where users interrupt the robot, label features, and provide demonstrations for the correct behavior. After training our framework, we will evaluate our algorithm in a final evaluation study. This section details our research questions, experimental setup, Institutional Review Board approved study design, and our model architecture.

### 3.1 Research Questions

**RQ1:** *Can we learn interpretable features from interventions?* We investigate whether we can learn to predict features of interest from participant interventions. We further validate whether the learned interpretable features generalize to novel users.

**RQ2:** *Does adding features improve performance over a baseline without features?* We evaluate whether understanding the relevant features improves the objective robot performance and the user's perception of the robot's performance.

**RQ3:** *Does communicating to users the anticipated feature of interest impact the users' perceptions of the robot?* We investigate whether communicating the predicted feature of interest when a user interrupts the robot changes the perceptions of the robot.

### 3.2 Experiment Setup

We use the Spot robot [1] to learn household chores via LfD.

*3.2.1 Domains.* We design four household tasks as the domains (see Figure 1) because prior work has shown that cleaning and chore tasks are relevant for assistive robots [35, 36]. Additionally, we chose these domains because chores are tasks that a robot could help with and each domain has human preferences (e.g., where to place a dish in the dishwasher). We add risk to some tasks (e.g., robot holding a knife) to increase the stakes for robot failures. We will further refine the design of the domains in a pilot study.

**Loading the Dishwasher:** The robot's goal is to place the plastic dish in the tabletop dishwasher. A preference could be having the dish placed upside-down.

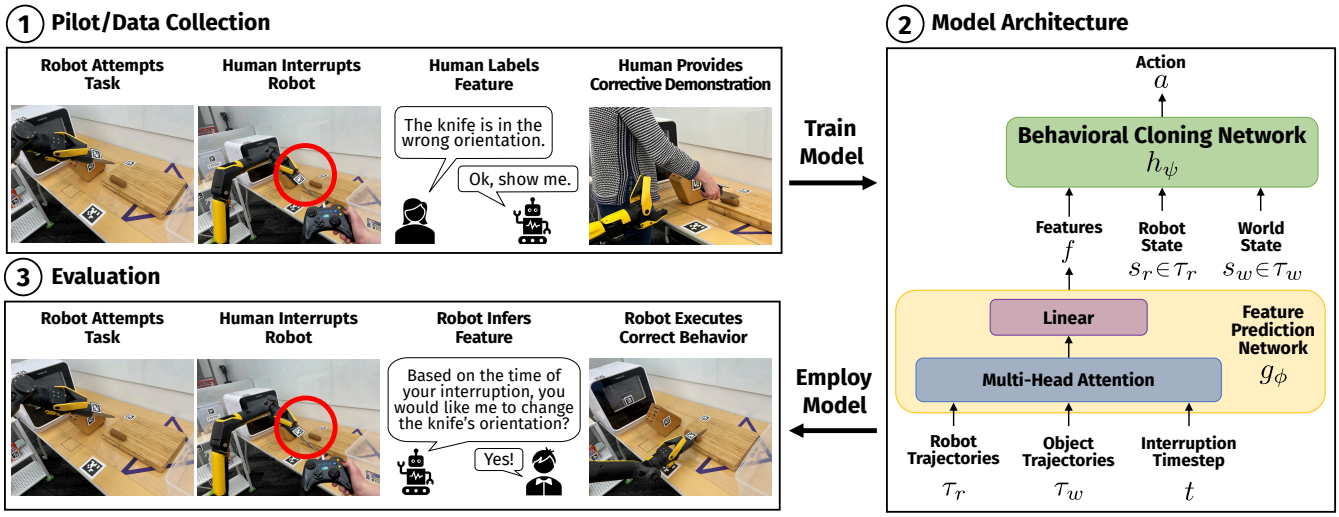


**Figure 2: (1) In the data collection study, the robot attempts the task, then the human interrupts the robot, by pressing the red button, because the robot is pointing the knife at the person. The human labels the feature as incorrect orientation, and then demonstrates the correct behavior to the robot. (2) We will use this data to train our framework. The behavioral cloning network, $h_\psi$, learns a robot action, $a$, from the robot state, $s_r$, state of the environment $s_w$, and predicted human feature set, $f$. These features, $f$, are outputs from the feature prediction network, $g_\theta$, which takes robot trajectory $\tau_r$, environment object trajectories $\tau_w$, and human intervention timestep, $t$. (3) During evaluation, after the human interrupts the robot, the robot predicts the feature of interest for that person. Then based on the feature, the robot executes the correct behavior.**

**Putting Bleach Away:** The robot's goal is to place the bleach bottle on the pantry shelf. A preference could be putting cleaning products on a different shelf than food items.

**Sweeping Up Glass:** With a hand-held broom, the robot's goal is to sweep the plastic "broken glass" into a dustpan without spilling. An example preference could be the broom's orientation.

**Cutting Food:** The robot must use a real knife to cut the play-dough cookie roll into properly-sized cookies to bake. A preference could be the angle at which the cookie dough is cut.

*3.2.2 Wizard-of-Oz Trajectories.* In each domain, we pre-specify six robot trajectories: two successes, two objective failures, and two subjective failures. To show the participants a consistent set of robot behaviors, the result of the robot's policy rollout is a pre-determined trajectory. An objective failure occurs when the robot fails to complete the task goal (e.g., colliding with the dishwasher). A subjective failure is when the robot achieves the goal without satisfying a user's preferences (e.g., placing the dish in the wrong orientation). We expect that for objective failures, people will interrupt uniformly, so we add subjective failures to study personalization.

## 3.3 Study Procedures

**Pilot Study** – We will run a pilot study to determine whether participants perceive each trajectory as intended (i.e., successful trajectories are perceived as successes, objective failures as failures, and subjective failures are *sometimes* perceived as failures). Past experiments have shown that participants do not intervene, even if the robot is colliding with objects [26]. As such, we will evaluate (and improve) our instructions to determine whether participants intervene when observing robot failures.

Participants will observe each trajectory in each domain. The order is based on the domain ordering condition (see Section 3.4). If a participant interrupts a trial, we will ask why they interrupted the robot. Their verbal explanation will be recorded via a microphone. After the study is concluded, these natural language explanations will be analyzed to determine a core set of important features across domains. We aim to collect a dataset of features of importance from the population rather than using an experimenter-defined dataset. Our goal is for these features to be generalizable to future domains.

To obtain these features, we first transcribe the interviews and conduct a thematic content analysis [3]. We will then design a graphic user interface (GUI) for users to choose which feature modifications are necessary post-interruption.

**Data Collection** – We propose to conduct a within-subjects ($n = 40$) data collection study where participants will observe the robot complete tasks in each of the domains, akin to the pilot study. After interrupting the robot, participants will use our developed GUI to indicate which feature the robot performed incorrectly. Then, participants will provide a demonstration via motion-capture, so we can understand how the robot should have performed the task differently. We will use this data to train our model (Section 3.5).

**Evaluation Study Design** – We will conduct a 4x4 between-subjects experiment with 16 participants per condition. The two factors of this study are the Domain Ordering and the Feature Conditions (Section 3.4). We will evaluate whether a robot policy that learns the important features characterizing an intervention is more effective at learning user preferences. Participants will first fill out pre-study surveys to collect demographic information. Then participants will observe the robot attempt a series of household tasks based on the Domain Ordering Condition. After each trial, participants will complete the post-trial surveys. If participants'

interrupt a trial, the robot will communicate the inferred feature or the user will input the feature using the GUI, dependent on the Feature Condition. The robot will then attempt the task again, with the policy updated based on the Feature Condition. Each participant will only experience one Feature Condition. After all trials, participants will complete the post-study surveys (Appendix 6).

### 3.4 Conditions

We describe the two types of conditions below:

**1) Domain Ordering Condition –** The participants experience four domains in this study. The order in which the domains are encountered is randomized and counterbalanced using Latin Squares to mitigate the learning effect (see Appendix Section 7.1).

In each domain the participant observes the robot execute a pre-determined policy for six trials, resulting in two successful outcomes and four failure outcomes (two objective and two subjective failures). Prior work has shown that the frequency and timing of robot errors impact user behavior and perception of the robot [12]. As such, the order of the trials (success vs. failure) for each sequential domain is randomly set and then held constant for all participants (see details in Appendix Section 7.2).

**2) Feature Condition –** In this study, the participant will observe the robot rollout its policy for multiple trials. The participant is asked to interrupt the robot when the robot makes a mistake. The robot will then infer the feature of importance and attempt the trial again. Feature Condition refers to how the inferred feature is employed. The feature conditions we consider are as follows.

- **Ours (learned feature):** The robot infers and communicates the feature (using templated language, e.g. "I think you interrupted because of X."). The robot then attempts the task using the learned feature (Section 3.5).
- **No feature:** As a baseline, the robot simply learns via BC.
- **Adversarial feature:** The robot will infer the feature, then choose a different feature to use and communicate to the user. This accounts for bias when working with adaptive systems.
- **Human chooses feature:** Similar to the data collection study, the user inputs the feature using the GUI. We want to determine if users prefer to tell the robot the feature or have the robot guess.

### 3.5 Model Architecture

In our model architecture (Figure 2.2), we learn the features via the Feature Prediction Network, $g_\phi$. The inputs to $g_\phi$ are the robot state trajectory, $\tau_r$, the states of the objects in the world, $\tau_w$, and the time of interruption, $t$. The network learns an embedding via an attention layer which is then fed through a linear layer. The output of $g_\phi$ is a classification encoding the feature of importance. We will train this model using cross-entropy loss and the labeled features from the data collection study. The robot will learn a trajectory policy using a behavioral cloning model as a baseline (no feature condition), where the robot state, $s_r$, and world object state, $s_w$, are mapped to the robot's action, $a$. In our framework, we add the feature vector, $f$, as an input to the policy network, $h_\psi$. The model assumes that the reason for interruption, given state, is homogeneous across people. Therefore, if two people interrupt in the same state, the model will predict the same feature. However, we assume that people will interrupt in different states based on their preferences.

### 3.6 Metrics

- **Framework Training Metrics –** We will first test our framework with a holdout dataset, assessing the accuracy of the feature predictions and model training time.
- **Pre-Study Metrics –** We will collect demographic information about participants, including personality [13] and participants' attitudes towards robots [29].
- **Post-Trial Metrics –** We will measure feature accuracy by asking participants if the robot guessed the correct feature. Objective accuracy of the robot's policy will be calculated based on task completion. To account for preferences, we will measure perceived robot accuracy on a success scale (1-10).
- **Post-Study Metrics** We will measure participants' reliance on the robot via intervention rate over all trials. We will compare participant responses on the Trust in Automation [22] subscales and assess usability [7], perceived safety [5], and workload [18].

## 4 RESULTS AND FUTURE WORK

We have completed our pilot study with 13 participants with a mean age of 23.8 and standard deviation of 1.03 (30.8% Female, 69.2% Male). We found that, on average, participants rated successful trials with a score of 7.8 out of 10, subjective failures with 5.8 and objective failures with 4.0. We aimed to design the tasks such that, successes will score 7-10, subjective failures 4-6, and objective failures 1-3. On the dishwasher task, participants rated successful trials lower than expected: 5.9, due to the robot releasing the dish from too high. On the bleach task, the objective failures were rated higher than expected: 5.6, due to not all participants rating collisions negatively. We plan to redesign the dishwasher task and tell participants that the robot should complete the tasks without colliding.

Additionally, many participants did not interrupt until after the robot failed irrecoverably (e.g., wiped all the glass on the floor). We plan to include a warn button, that does not stop the robot, so participants can indicate when the robot might be about to make a mistake. This way, the human corrective demonstrations can show how to avoid failure instead of starting after the interruption point.

From the pilot study, the features that participants provided are orientation of the object, position of the object relative to other objects in the environment, and speed of the robot. We will incorporate these features into the GUI. Next, we plan to conduct the data collection study, train the model, and then evaluate our framework with the evaluation study. In the future, we also aim to conduct this study with a target population of older adults. Our goal is to develop a framework that can generalize to new users and enable them to personalize robot behavior through simple interventions.

## 5 CONCLUSION

We propose a multi-phase study to learn interpretable features from interventions. We posit that we can learn people's preferences for robot behavior based on when someone decides to stop a robot. We design an experiment where the robot performs household chores with varying levels of success and participants stop the robot when it is making a mistake. We plan to collect data from users to train a framework to learn features that inform a robot's behavior policy.

# REFERENCES


[1] Feb 2024. URL https://bostondynamics.com/products/spot/.

[2] Pieter Abbeel and Andrew Y. Ng. Apprenticeship Learning via Inverse Reinforcement Learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1-58113-838-5. doi: 10.1145/1015330.1015430. URL https://doi.org/10.1145/1015330.1015430. event-place: Banff, Alberta, Canada.

[3] Rosemarie Anderson. Thematic content analysis (tca). *Descriptive presentation of qualitative data*, 3:1–4, 2007.

[4] Andrea Bajcsy, Dylan P. Losey, Marcia K. O'Malley, and Anca D. Dragan. Learning from Physical Human Corrections, One Feature at a Time. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, pages 141–149, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-4953-6. doi: 10.1145/3171221.3171267. URL https://doi.org/10.1145/3171221.3171267. event-place: Chicago, IL, USA.

[5] Christoph Bartneck, Elizabeth Croft, and Dana Kulic. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009. doi: 10.1007/s12369-008-0001-3.

[6] Andreea Bobu, Marius Wiggert, Claire Tomlin, and Anca D Dragan. Feature expansive reward learning: Rethinking human input. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 216–224, 2021.

[7] John Brooke. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996. Publisher: London, England.

[8] Guoting Chang and Dana Kulić. Robot task error recovery using Petri nets learned from demonstration. In *2013 16th International Conference on Advanced Robotics (ICAR)*, pages 1–6, 2013. doi: 10.1109/ICAR.2013.6766465.

[9] Justin Cheng and Michael S Bernstein. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 600–611, 2015.

[10] Sonia Chernova and Manuela M. Veloso. Interactive Policy Learning through Confidence-Based Autonomy. *Journal of Artificial Intelligence Research*, 34:1–25, 2009. doi: doi:10.1613/jair.2584.

[11] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 351–360, 2021.

[12] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 251–258, 2013. doi: 10.1109/HRI.2013.6483596.

[13] M Brent Donnellan, Frederick L Oswald, Brendan M Baird, and Richard E Lucas. The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192, 2006.

[14] Boi Faltings, Pearl Pu, and Paolo Viappiani. Preference-based Search using Example-Critiquing with Suggestions. *The journal of artificial intelligence research*, 27:465–503, December 2006. doi: 10.1613/jair.2075.

[15] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy Shaping: Integrating Human Feedback with Reinforcement Learning. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/e034fb6b66aacc1d48f445ddfb08da98-Paper.pdf.

[16] Shuai Han, Mehdi Dastani, and Shihan Wang. Sample efficient reinforcement learning by automatically learning to compose subtasks, 2024.

[17] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y.C. Chen, Ewart J. De Visser, and Raja Parasuraman. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, 53(5):517–527, 2011. ISSN 00187208. doi: 10.1177/0018720811417254. ISBN: 0018720811417.

[18] S. G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. 1988.

[19] Jason Johnson. *Type of Automation Failure: The Effects on Trust and Reliance in Automation*. PhD Thesis, 2004. Issue: December.

[20] Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. *HG-DAgger: Interactive Imitation Learning with Human Experts*. 2018. _eprint: 1810.02890.

[21] Bradley Knox and Peter Stone. TAMER: Training an Agent Manually via Evaluative Reinforcement. In *2008 7th IEEE International Conference on Development and Learning*, pages 292–297, 2008. doi: 10.1109/DEVLRN.2008.4640845.

[22] Moritz Körber. Theoretical considerations and development of a questionnaire to measure trust in automation. March 2018.

[23] Jonathan Lee, Michael Laskey, Roy Fox, and Ken Goldberg. Constraint Estimation and Derivative-Free Recovery for Robot Learning from Demonstrations. In *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, pages 270–277, 2018. doi: 10.1109/COASE.2018.8560342.

[24] Meredith Mealer, Ellen L. Burnham, Colleen J. Goode, Barbara Rothbaum, and Marc Moss. The prevalence and impact of post traumatic stress disorder and burnout syndrome in nurses. *Depression and anxiety*, 26(12):1118–1126, 2009. ISSN 1520-6394 1091-4269. doi: 10.1002/da.20631. Place: United States.

[25] Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. Should robots be obedient?, 2017.

[26] Nina Moorman, Erin Hedlund-Botti, Mariah Schrum, Manisha Natarajan, and Matthew C. Gombolay. Impacts of Robot Learning on User Attitude and Behavior. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, pages 534–543, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 978-1-4503-9964-7. doi: 10.1145/3568162.3576996. URL https://doi.org/10.1145/3568162.3576996. event-place: Stockholm, Sweden.

[27] Carl Mueller, Jeff Venicx, and Bradley Hayes. Robust Robot Learning from Demonstration and Skill Repair Using Conceptual Constraints. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6029–6036, 2018. doi: 10.1109/IROS.2018.8594133.

[28] Scott Niekum, Sachin Chitta, Andrew Barto, Bhaskara Marthi, and Sarah Osentoski. Incremental Semantically Grounded Learning from Demonstration. In *Robotics: Science and Systems (R:SS)*, June 2013. doi: 10.15607/RSS.2013.IX.048.

[29] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. Measurement of negative attitudes toward robots. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 7(3):437–454, 2006.

[30] Harish Ravichandar, Athanasios S. Polydoros, Sonia Chernova, and Aude Billard. Recent Advances in Robot Learning from Demonstration. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):297–330, May 2020. ISSN 2573-5144. doi: 10.1146/annurev-control-100819-063206. URL https://doi.org/10.1146/annurev-control-100819-063206. Publisher: Annual Reviews.

[31] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[32] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. No-Regret Reductions for Imitation Learning and Structured Prediction. In *14th Internation Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, 2011. URL https://arxiv.org/abs/1011.0686v3.

[33] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://doi.org/10.1038/s42256-019-0048-x.

[34] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *ACM/IEEE International Conference on Human-Robot Interaction*, 2015-March:141–148, 2015. ISSN 21672148. doi: 10.1145/2696454.2696497. ISBN: 9781450328821 Publisher: ACM.

[35] Lakshmi Seelam, Erin Hedlund-Botti, Chuxuan Yang, and Matthew Gombolay. Interface Design for Learning from Demonstration with Older Adults. In *Association for the Advancement of Artificial Intelligence Fall Symposium Series*, 2023.

[36] Cory-Ann Smarr, Tracy L. Mitzner, Jenay M. Beer, Akanksha Prakash, Tiffany L. Chen, Charles C. Kemp, and Wendy A. Rogers. Domestic Robots for Older Adults: Attitudes, Preferences, and Potential. *International journal of social robotics*, 6(2): 229–247, April 2014. ISSN 1875-4791 1875-4805. doi: 10.1007/s12369-013-0220-0.

[37] Jonathan Spencer, Sanjiban Choudhury, Matt Barnes, Matthew Schmittle, Mung Chiang, Peter Ramadge, and Siddhartha Srinivasa. Learning from Interventions: Human-robot interaction as both explicit and implicit feedback. 2020. doi: 10.15607/RSS.2020.XVI.055.

[38] Adriana Tapus, Maja J. Mataric, and Brian Scassellati. Socially assistive robotics [Grand Challenges of Robotics]. *IEEE Robotics & Automation Magazine*, 14(1): 35–42, 2007. doi: 10.1109/MRA.2007.339605.

[39] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon, and Myrthe L Tielman. Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*, pages 3–12, 2020.

[40] Alexandra Zytek, Ignacio Arnaldo, Dongyu Liu, Laure Berti-Equille, and Kalyan Veeramachaneni. The need for interpretable features: motivation and taxonomy. *ACM SIGKDD Explorations Newsletter*, 24(1):1–13, 2022.

# 6 APPENDIX: INTERVIEW QUESTIONS

**Post-Study Interview Questions**

- What can the robot do?
- What can't the robot do?
- What does the robot know/ understand?
- What does the robot not know/ understand?
- When did you interrupt the robot?
- What were your thoughts on the process of working with the robot?
- Do you have any feedback for us about the user interface?
- Was there anything missing from the interface?

**Post-Interruption Interview Questions**

- Why did you interrupt the robot?
- What did the robot do wrong?
- How should the robot's behavior change?
- Describe how the robot should do this task.
- Do you trust the robot?
- Do you trust the robot to do this task?

**Post-Rollout Interview Questions**

- On a scale of 1 to 10, was the robot successful in this trial?

# 7 APPENDIX: CONDITIONS

## 7.1 Domain Ordering Condition

The following table lists the domain ordering in the four domain ordering conditions, obtained via a Latin square. Each participant experiences one domain ordering condition.

| | **Domain 1** | **Domain 2** | **Domain 3** | **Domain 4** |
|---|---|---|---|---|
| **Domain Ordering 1** | Filling the Dishwasher | Putting Bleach Away | Sweeping Up Glass | Cutting Food |
| **Domain Ordering 2** | Putting Bleach Away | Sweeping Up Glass | Cutting Food | Filling the Dishwasher |
| **Domain Ordering 3** | Sweeping Up Glass | Cutting Food | Filling the Dishwasher | Putting Bleach Away |
| **Domain Ordering 4** | Cutting Food | Filling the Dishwasher | Putting Bleach Away | Sweeping Up Glass |

## 7.2 Outcome Ordering in Each Domain

The following table lists the ordering of outcomes in the domains the participant experiences sequentially, randomized then held constant for each participant.

| | **Domain 1** | **Domain 2** | **Domain 3** | **Domain 4** |
|---|---|---|---|---|
| **Outcome 1** | Objective Failure | Objective Failure | Objective Failure | Success |
| **Outcome 2** | Subjective Failure | Success | Subjective Failure | Subjective Failure |
| **Outcome 3** | Subjective Failure | Success | Success | Success |
| **Outcome 4** | Objective Failure | Subjective Failure | Subjective Failure | Objective Failure |
| **Outcome 5** | Success | Subjective Failure | Success | Objective Failure |
| **Outcome 6** | Success | Objective Failure | Objective Failure | Subjective Failure |