DATA-DRIVEN PERSONALIZATION TECHNIQUES TO ACCOUNT FOR HETEROGENEITY IN HUMAN-MACHINE INTERACTION

A Dissertation
Presented to
The Academic Faculty

By

Mariah Schrum

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mechanical Engineering
Institute for Robotics and Intelligent Machines

Georgia Institute of Technology

May 2023

DATA-DRIVEN PERSONALIZATION TECHNIQUES TO ACCOUNT FOR HETEROGENEITY IN HUMAN-MACHINE INTERACTION

Thesis committee:

Dr. Matthew Gombolay
School of Interactive Computing
Georgia Institute of Technology

Dr. Bill Smart
College of Manufacturing, Industrial, and Manufacturing Engineering
Oregon State University

Dr. Sonia Chernova
School of Interactive Computing
Georgia Institute of Technology

Dean Ayanna Howard
College of Engineering
Ohio State University

Dr. Karen Feigh
Department of Aerospace Engineering
Georgia Institute of Technology

Date approved: February 28, 2023

I may not have gone where I intended to go, but I think I have ended up where I needed to be.

Douglas Adams

For my grandparents who enabled me to pursue higher education.

ACKNOWLEDGMENTS

I would like to take the opportunity to acknowledge the contributions of those who have supported me in my journey to obtaining a doctorate of philosophy in robotics. First and foremost, I would like to thank my advisor, Dr. Matthew Gombolay. As a first year PhD student, I had little to no knowledge or background in machine learning or human-robot interaction and was uncertain about the direction I wanted to go in for my PhD. Despite this, Matthew accepted me as a member of his lab. Throughout the five years of my PhD, Matthew provided me with the guidance necessary to mold me into a successful researcher in algorithmic HRI. Not only is Matthew a supportive advisor, but he also created an incredible research community to operate within. The culture of a lab can have an immense impact on one's PhD experience, and I could not have asked for a better community in which to obtain my PhD than the CORE Robotics Lab. Matthew fostered both a diverse and collaborative environment that made my PhD experience that much better. I am incredibly grateful that he took a chance with me as one of his first students and I cannot thank him enough for his guidance and mentorship.

I am honored to have been supported by an incredible committee. Dr. Sonia Chernova always made the time to give me advice and has been a fantastic role model for me and the other women in the robotics PhD program. Dr. Ayanna Howard mentored me during my first semester at Georgia Tech and during my time in the ARMS fellowship program. Her advice and support during the early stages of my PhD put me on a track to become a successful PhD student. I am grateful for Dr. Karen Feigh's advice on human factors research and her informative and constructive feedback. She made me a better researcher. Dr. Bill Smart joined my committee despite not knowing me beforehand and became a valuable and supportive mentor.

My journey as a researcher started at my undergraduate institution, Johns Hopkins. During my time at Johns Hopkins in the Whiting School of Engineering, I worked with Dr.

Russ Taylor and it is thanks to him that I started on the path to my PhD. Dr. Taylor is an incredible researcher and scientist and it was a privelege to work with someone who has done so much to advance the field of surgical robots. Dr. Taylor provided me with many research opportunities that few undergraduates have the fortune to experience. Although I was only an undergraduate, Dr. Taylor gave me my own research projects and found the time in his busy schedule to meet with me weekly and provide guidance on my projects. My experience working with Dr. Taylor had an immense impact on me as a young researcher and I am incredibly grateful for his mentorship and for how he shaped my career path. I also want to thank everyone else in the CIIS lab who welcomed me into the lab and opened my eyes to the world of robotics research.

Part of what made my PhD fulfilling is the amazing collaborations that I was able to take part in. I want to thank all of my co-authors and collaborators for the part they played in my PhD. My collaborators at Emory, Dr. Sam Broida, Dr. Eric Yoon, Dr. Sangwook Yoon, Eric Cole, Dr. Mark Connolly and Dr. Robert Gross, thank you for your invaluable insight and medical knowledge. The exciting work in diagnosing lumbar spine disorders and optimizing deep brain stimulation would not have been possible without your efforts. I'd also like to thank the talented undergraduates and graduate students that I had the pleasure to work with including Nathaniel Belles, John Zhang, Wendell Hom, Tanner Beard, Steve Zakhorov, and Karthik Shaji. Despite the many UAVs we crashed, you kept working hard. Your dedication and perseverance inspire me.

I would be remiss if I didn't thank the hundreds of people, many of whom were Georgia Tech students, who participated in my many studies. This thesis would be worthless without your valuable time and data.

I am so privileged to have been a part of a large, diverse lab and I would like to thank all of my lab mates in the CORE Robotics Lab. I tell everyone that our lab has the best lab culture. The incredibly intelligent, supportive, and kind people that comprise the lab are the reason why. I particularly want to thank my cohort Esi Seraj and Rohan Paleja for being

there since day one. Esi, thank you for your controls expertise, friendship, and the great times we had in Japan and New Zealand. Rohan, thanks for all of the awesome memories we've made since riding together to Georgia Tech's campus from the Atlanta airport back in 2018. You both made my PhD experience inside and outside of the lab a hundred times better and there is no one else I would rather stumble through a PhD with than the two of you. I also want to thank my office mates, Erin Hedlund-Botti, Pradyumna Tambwekar, Andrew Silva, and Nina Moorman. Erin, you were a constant sounding board for so many of my ideas and a great collaborator and coauthor. Thank you also for being a good friend. Prad, thank you for always being there to talk through any problem with me. Whether it was technical, academic, or personal, you always found time to listen and give me advice. Andrew, thank you for your machine learning expertise, for making me a better researcher, and for our summer internship adventures. I am still exhausted from our Mount Tallac hike. Nina, it has been awesome to work with you in the early stages of your PhD and watch you grow as a researcher. Thanks for all of the memories in Greece! Thank you to everyone else in the CORE Robotics lab, including Manisha Natarajan, Zac Chen, Sam Yi Ting, Sean Ye, Josh Bishop, Laura Strickland, Zulfiqar Zaidi, and Dr. Nakul Gopalan for your endless encouragement, support, and friendship.

During my PhD, I had the opportunity to intern at two amazing companies: Intuitive Surgical and Toyota Research Institute. During these internships, I worked with several amazing teams of researchers and many incredible mentors. I especially want to thank Kevin Pluckter, Dr. Troy Adebar, and the rest of the team at Intuitive Surgical. You all made my internship experience invaluable and I was humbled to have the opportunity to join the incredible mission at Intuitive. Thank you to everyone I had the pleasure of working with at Toyota Research Institute and in particular, my mentor, Dr. Andrew Best for your guidance and mentorship during my internship. I learned so much from you, and working with the simulator that you and your team designed and built was an amazing experience.

On a more personal note, I want to thank all of my friends that I have made at Georgia

Tech. Michael Johnson, Andrew Messing, Leng Ghuy, and Glen Neville, your friendship outside of the lab has meant the world to me. Our trips to Asheville, Chattanooga, Nashville, Panama City, Hilton Head, and Seattle, to name just a few, are some of my best memories. Arielle Berman and Divya Srivastava thank you both for your love and friendship and for being there to grab a drink and talk. I want to thank the members of the Volley Llamas, Michael Johnson, Ben Shaffer, Jenny Leestma, Andrew Messing, Aakash Bajpai, Ethan Schonhaut, Rohan Paleja, Reese Peterson, and Jordyn Schroeder for their support both on and off the court. The stress relief of playing volleyball with good friends got me through some tough moments in my PhD. Thank you to Sarai Sherfield, Patrick Grady, Carter Price, Gerry Chen, Jack Kolb, Bruce Wingo, Maithili Patel, and everyone else who took the time to travel with me, hang out at conferences, de-stress at parties, and overall made the PhD journey much more enjoyable. Outside of the world of academia, I am so grateful for the friendship of Tara Napolitano, Lauren Clark, and Laura Andreola over these last 15+ years. My friends have made my PhD experience fun and I feel so privileged to have shared this experience with you all.

Finally, I would not be where I am if not for the support of my family. Thank you Lynn Schrum, Sid Schrum, and Mailande Schrum for your constant support, listening to my defense practice, and always asking insightful questions. Thank you Dan Schrum for being a great brother and travel companion. Lastly, Mom and Dad, thank you for believing in my abilities and convincing me that I am capable of earning a PhD. Thank you for answering the many stress induced phone calls and encouraging me when I felt lost or incapable. I would not have been able to achieve this goal without your support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

a/d anticipatory-/delayed-

ADB Aggressive Driving Behavior

ANOVA Analysis of Variance

AV Autonomous Vehicle

BC Behavioral Cloning

DAgger Dataset Aggregation

DBS Deep Brain Stimulation

DTW Dynamic Time Warping

EI Expected Improvement

EIL Expert Intervention Learning

EPN Embedding Predictor Network

HG-DAgger Human-Gated Dataset Aggregation

IRB Institutional Review Board

IRL Inverse Reinforcement Learning

LAL Learning Active Learning

LfD Learning from Demonstration

LSTM Long-Short Term Memory Network

MAVERIC Manipulating Autonomous Vehicle Embedding Region for Individuals' Comfort

MDP Markov Decision Process

Meta BO Meta-Bayesian Optimization

MILP Mixed-Integer Linear Program

MIND MELD Mutual Information Driven Meta-Learning from Demonstration

MPC Model Predictive Control

MSE Mean-squared error

NASA TLX NASA Task Load Index

o/u over-/under-

PI Proportional and Integral

POMDP Partially Observable Markov Decision Process

RC Robot-Centric

RRT Rapidly-exploring Random Trees

Safe MetAL Safe Meta-Active Learning

SVM Support Vector Machine

TRI Toyota Research Institute

Summary

As robots and AI systems become more prevalent in every-day life, humans and machines will have to work closely together. Robotic devices will be used to support human health, service robots will operate alongside humans in homes, and autonomous vehicles will have to safely drive end-users to their destination. Yet, humans exhibit a high degree of heterogeneity which poses a challenge for robotic systems that are tasked with learning from and supporting humans. For example, in a medical setting, individual patients are likely to have different needs and varying biology that must be accounted for. Autonomous Vehicles (AVs) will have to learn about the differing preferences of end-users and adapt accordingly. Because of this human heterogeneity, one-size-fits-all algorithms will not suffice in many human-machine interaction scenarios. Instead, to effectively support humans, machines must be capable of recognizing individual desires, abilities, and characteristics and adapt to account for differences across individuals. This thesis focuses on the development of personalized algorithms that enable machines to better support and work with humans. Specifically, I aim to develop and research novel techniques for safely and efficiently supporting heterogeneous humans across various robotic domains. In this work, I develop data-driven, personalized frameworks in healthcare, learning from demonstration, and autonomous driving domains to account for heterogeneity amongst end-users.

In this thesis, I first investigate the question of how robots can best learn from human demonstrators who are suboptimal and heterogeneous in their suboptimality. I present my work on Mutual Information Driven Meta-Learning from Demonstration (MIND MELD), a framework enabling personalized robotic learning from heterogeneous human demonstrators via a learned, personalized embedding. I then extend this approach with Reciprocal MIND MELD and introduce a framework to provide personalized feedback to suboptimal human demonstrators to improve upon their ability to provide high quality demonstrations. Humans are not only heterogeneous in terms of their abilities when teaching machines; they also

tend to differ in their preferences for various machine behaviors. To account for differing preferences amongst end-users, I draw upon our Reciprocal MIND MELD work and introduce Manipulating Autonomous Vehicle Embedding Region for Individuals' Comfort (MAVERIC), an approach for personalizing driving styles of AVs to fit the preferences of end-users via personalized embeddings. In my final work, I consider personalization in safety critical domains such as healthcare. I introduce Safe Meta-Active Learning (Safe MetAL), an approach for determining the optimal, personalized parameter settings for a Deep Brain Stimulation (DBS) patient.

The contributions of this thesis are as follows:

- **Creation of a novel meta-learning architecture for learning from heterogeneous, non-expert demonstrators (HRI '22)** [1, 2]: I introduce Mutual Information-driven Meta-learning from Demonstration (MIND MELD), which meta-learns a person-specific mapping from human-provided, corrective labels to idealized labels in a robot-centric learning from demonstration paradigm.

- **Development of a framework for teaching heterogeneous humans to be better demonstrators via personalized feedback (CoRL '22)** [3]: I propose Reciprocal MIND MELD, which builds upon MIND MELD and provides verbal feedback to provide personalized corrections for teacher suboptimality based upon the learned personalized embedding, thereby improving the teacher's ability to provide high quality demonstrations.

- **Creation of a data-driven framework for personalized autonomous driving** [4]: I develop Manipulating Autonomous Vehicle Embedding Region for Individuals' Comfort (MAVERIC) in which we learn to mimic an end-user's driving style via a learned personalized embedding and investigate the factors impacting the effect of homophily to optimize autonomous vehicle driving style.

- **Development of a meta-active learning framework for efficient and safe person-**

alization in a healthcare setting (IROS '22) [5]: I introduce Safe MetAL (Safe Meta-Active Learning), a hybrid meta-learning and mathematical programming approach that enables efficient, safe, and computationally fast learning of the optimal parameter settings for a DBS patient's brain

CHAPTER 1
INTRODUCTION

1.1 Motivation

A fundamental aspect to the survival of the human species is human diversity [6]. Diversity is positively correlated with survival due to a number of factors. Greater biodiversity within a species leads to greater stability due to increased robustness to disease, climate changes, and other disturbances [7]. Diversity engenders specialization and the division of labor which improves the population's resiliency and sustainability. While genetics play a large role in diversity, plasticity and the ability to change in response to the environment also fosters heterogeneity within a population [8]. The human brain in particular has a large plastic potential which is one of the main drivers behind the high degree of heterogeneity amongst humans beings. Genetics lay the foundation for diversity at birth whereas plasticity promotes increased diversity during development and throughout the lifespan [9].

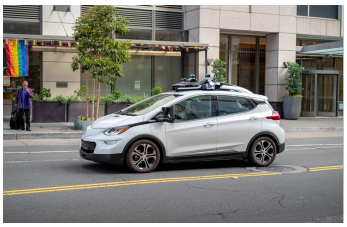
This diversity, so crucial to the survival of the human species, poses a unique challenge to machines that are designed to interact with human end-users. The high degree of heterogeneity amongst humans means that each end-user is a unique system that the machine must learn about and adapt to so as to optimize the human-machine relationship. There are many latent variables that govern human preferences, behavior, and decision making that machines must take into consideration. For example, an individual's personality and prior experiences may impact their level of trust in an autonomous system. Biological differences stemming from both genetics and environment influence how an AI system can best respond to and support a patient in a healthcare setting [10]. In a human-machine team, an individual's level of skill at accomplishing a task is often an unknown variable. To optimize the human-machine team, the machine must learn about the skill of its human

partner so as to adjust its behavior to benefit the team.

Because of the differences amongst individuals, one-size-fits-all approaches are not the best solution for optimizing human-machine interaction. Instead, personalization is required to promote engagement, improve the overall perception of the machine, and to optimize the human-machine relationship [11]. To optimize their behavior for an individual end-user, human-machine systems must be capable of personalizing and dynamically adapting to the specific end-user they are working with so as to account for individual differences. There are two methods by which a machine can adapt to an individual end-user: the machine can either be *adaptable* or *adaptive* [12]. Adaptable systems provide the end-user with the means to alter the system to meet their needs. Adaptable systems require less a priori knowledge about the end-user and therefore can be easier to design and implement. However, in such a paradigm, the burden falls to the end-user to actively change the behavior of the system to fit their preferences and needs. This additional requirement can pose challenges due to the increased workload, general inconvenience, and the inorganic nature of the interaction.

Adaptive systems on the other hand actively learn about, create a model of, and autonomously adjust to the end-user's needs and preferences. Because this dynamic adaptation results in more organic and less burdensome interactions, prior work has shown that adaptive systems are viewed as more competent and are more relied on by end-users, [12]. The downside is that these systems can be more prone to failure due to lack of data, low-quality data, or ineffective learning algorithms. However, because prior work has shown that end-users prefer adaptive systems, in my thesis, I propose that we should strive to create robust, data-driven approaches capable of actively and dynamically personalizing to the needs and preferences of end-users.

As shown in Figure 1.1 there are many applications in which personalization of human-machine systems for individuals is necessary and each application requires the machines to reason about different aspects of the end-user and to optimize for different objectives. When an individual purchases a new robot and brings it to their home, they will want the



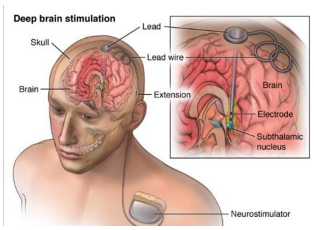


Figure 1.1: This figure shows several domains (i.e., autonomous driving, learning from demonstration, and healthcare) in which personalization of the machine is necessary to optimize the relationship with the human. (Courtesy: https://www.wired.com/story/news-rules-clear-way-self-driving-cars/, http://www.cairo-lab.com/papers/hri20m.pdf, https://www.mdpi.com/2079-9292/11/6/939)

robot to accomplish tasks with unique objectives and challenges specific to their needs and environtments [13]. To learn how to accomplish these novel tasks, robots can utilize Learning from Demonstration (LfD) techniques. In LfD, the robot learns a policy that maps the state of the world to how the robot should act to accomplish the human-specified or demonstrated task [14]. LfD techniques allow a non-expert, non-roboticist to teach a robot without the need for programming experience or robotics knowledge [14]. Yet, when teaching a robot, humans often exhibit various demonstrations styles, preferences, and suboptimalities [15, 16]. Therefore, robots will not only have to be capable of learning unique tasks to suit the needs of individuals but will also have to take into account differing demonstration styles and suboptimal tendencies so as to optimally learn from heterogeneous demonstrators.

Autonomous Vehicles (AVs) are another prominent example of an autonomous system that must adapt to the end-user. Humans exhibit diverse driving styles and naturally will want their AV to display a driving style that matches their expectations and preferences. Alternatively, in healthcare domains, human-machine systems must consider the individual's biology and adapt to the various needs of the individual patient while also ensuring patient safety. In each domain, the machine must consider different characteristics about the individual and account for various driving styles.

For systems to be capable of autonomously adapting to individual end-users, these

systems will need to gather data about the end-user to provide insight into the user's personality, biology, preferences, and other relevant characteristics. This data can be gathered either by actively probing the end-user for information or passively observing the end-user's behavior and interactions. The downside to personalized approaches is that gathering enough data to train individual models for each human end-user is costly and often not practical. Constantly querying an end-user for information is taxing and burdensome to the end-user and passively observing the end-user may not provide adequate information [17]. Fortunately, while humans are heterogeneous in many respects, they also tend to share similarities with other humans in the population. For instance, while personality differs amongst individuals, personality falls along a spectrum and therefore, an individual's proximity to others offers rich information about the individual in question. The way in which one highly conscientious individual operates in the world will likely share many similarities with others high in trait conscientiousness.

Human-machine systems that consider the individual end-user as an isolated entity without viewing them within the context of the larger population neglect important information that may be important for determining the optimal behavior. When determining how best to interact with an end-user, human-machine systems should instead capitalize on data from both the individual in question as well as the relevant population. Systems that do so will have a greater ability to optimize for individual end-users by gathering salient information from both the individual and the context of the individual within the population.

Based upon insights from prior work, in this thesis, I identify three key challenges that must be considered when creating personalized, autonomous systems. First, these systems must be capable of gathering sufficient data so as to effectively learn about an individual and personalize appropriately. Additionally, they ideally should do so in a minimally invasive manner. I propose that, to gather a sufficient amount of data, systems should capitalize on both information derived from the population as well as the individual in question. Instead of learning from scratch with each new individual, systems should utilize prior information

derived from previous interactions in a similar scenario and from previous end-users to inform their personalized model of the target end-user. By doing so, machines will be able to effectively learn and adapt to an individual while reducing the number of queries that the system must make of the end-user.

Second, machines must be capable of personalizing to end-user in diverse domains. As discussed above there are many different scenarios and situations in which personalization is necessary and each of these domains requires different considerations and optimization objectives. For example, in a human-machine team, the machine should be personalized to account for the human's suboptimality with respect to the task objective. However, in the case of a fully autonomous vehicle, the way in which the human is suboptimal with respect to driving is not important; instead the individual's preferences for different driving styles or routes should be optimized.

Third, in safety-critical domains, machines must consider the safety of the end-user when providing personalized support. Healthcare is an example of a domain in which both safety and personalization are critical considerations. Machines and AI systems designed to support patients must be equipped with additional information, typically provided by domain experts, about how the machine's behavior may affect the safety of the patient. For example, in the domain of DBS, probing for information about the optimal parameter settings may cause unwanted side effects for the patient [18]. Therefore, domains such as healthcare require AI systems to reason about two often conflicting objectives - both maximizing benefits of treatment while also reducing potential side effects or dangerous outcomes.

In my work, I aim to both provide human-machine systems with the ability to personalize to fit the needs and preferences of end-users across many domains while simultaneously rigorously validating these methods in large human subject studies. Much of the prior work in personalization focuses on only one of these aspects. Prior work has either employed Wizard-of-Oz studies to determine end-users' reaction to a certain robotic behavior without actually

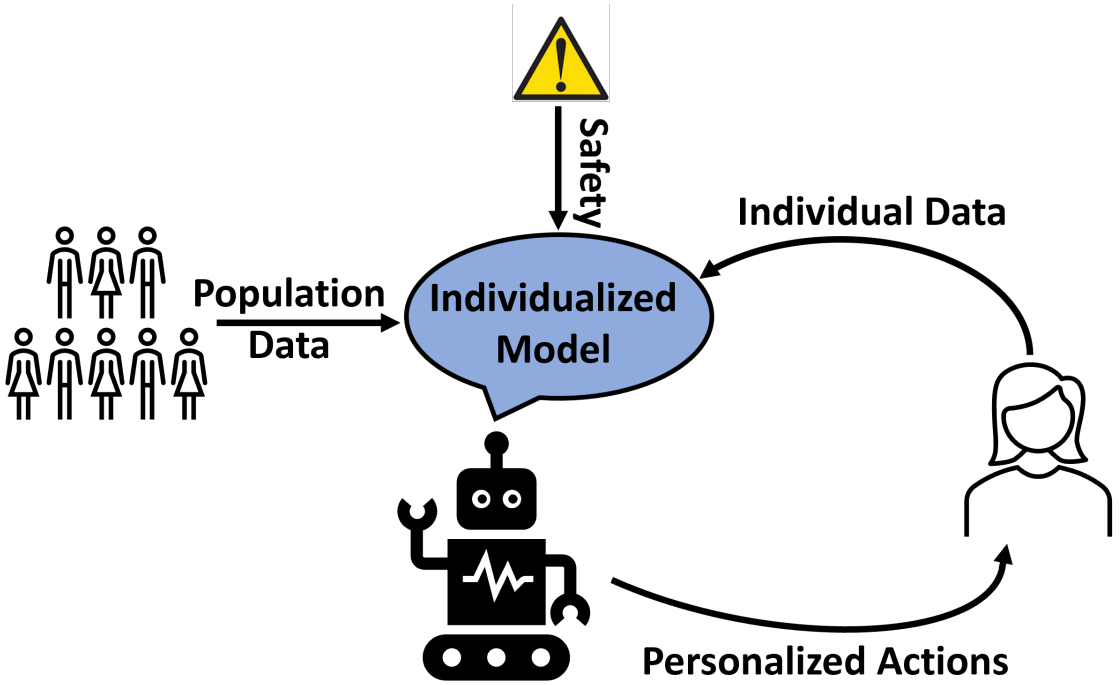


Figure 1.2: This figure shows an overview of my thesis. In my thesis I aim to develop algorithms to provide autonomous systems with the ability to learn from both the population and the individual to inform personalized interactions while also ensuring safety of the end-user.

creating a framework to produce this behavior [19, 20, 21] or has proposed frameworks to produce specific behaviors without validating the approach in a large human subjects study [22, 23, 24]. While there are examples of prior work that do both, the field lacks algorithmic approaches for human-machines interaction which have been validated in large human-subject studies. In my work, I aim to fill this gap and develop algorithms to produce personalized behavior while also rigorously validating these approaches with large human subject experiments.

Figure 1.2 shows an overview of my thesis in which I enable machines to 1) learn from both the population as well as the individual to effectively personalize their behavior, 2) do so in a variety of domains with diverse objectives, and 3) consider safety of the end-user. In my thesis, I additionally aim to rigorously validate the novel approaches I develop in large human subject studies.

1.2 Thesis Statement and Contributions

I present my thesis statement: Personalized algorithms that account for heterogeneity amongst individuals will improve upon the human-machine relationship across many domains including healthcare, autonomous vehicles, and learning from demonstration paradigms.

To support this claim, my work makes the following contributions summarized in Subsection 1.3.1-Subsection 1.3.4.

1.3 Outline of Dissertation Document

This thesis is organized as follows. Chapter 2 provides background on prior work and relevant approaches in personalization for human-machine system as related to my contributions discussed in Section 1.2. Chapter 3 and Chapter 4 focus on my frameworks Mutual Information Driven Meta-Learning from Demonstration (MIND MELD) and Reciprocal MIND MELD for personalized learning from suboptimal demonstrators. Chapter 5 discusses an extension of this approach for personalization of AV driving styles called MAVERIC. In Chapter 6, I seek to impose safety constraints when adapting to end-users in safety critical domains via our approach called Safe MetAL. I then discuss psychometric-based recommendation for best practices for human subject studies that have informed my research in Chapter 7 and provide concluding remarks in Chapter 8. Lastly, I discuss directions for future research in Chapter 9.

In Subsection 1.3.1-Subsection 1.3.4, I provided an overview of each of the novel contributions of my thesis.

1.3.1 Mutual Information Driven Meta-Learning from Demonstration

For robots to perform novel tasks in the real-world, they must be capable of learning from heterogeneous, non-expert human teachers across various domains. This capability enables

robot to personalize their behavior to fit the needs of users in diverse environments and learn new skills and tasks on-the-fly. Yet, novice human teachers often provide suboptimal demonstrations, making it difficult for robots to successfully learn [25]. Furthermore, humans tend to be heterogenous in the way in which they are suboptimal, making a one-size-fits-all approach impractical. Therefore, to effectively learn from humans, we must develop personalized learning methods that can account for teacher suboptimality and can do so across various robotic platforms.

To successfully learn from heterogeneous and suboptimal demonstrators, I propose to learn the "suboptimal style" (i.e., suboptimal tendency) by which an individual provides corrective feedback. Using this information, we map suboptimal, human feedback to feedback closer to optimal, thus improving the ability of the robot to learn from human demonstrators. I call this approach Mutual Information Driven Meta-Learning from Demonstration (MIND MELD) [2]. MIND MELD meta-learns a personalized embedding describing an individual's feedback style and maps suboptimal feedback to better feedback, conditioned on this learned style. I demonstrate MIND MELD's ability to outperform prior work in both human- and robot-centric LfD in a human subjects study.

Approach: To learn about the way in which an individual demonstrator is suboptimal in providing demonstrations, I design calibration tasks. These tasks are Wizard-of-Oz [26] rollouts representative of a policy learned via LfD and are drawn from the distribution of possible tasks that the demonstrator may encounter [27]. I then calculate the optimal, corrective feedback (i.e., steering angle of the car) for each point in time along each calibration trajectory. Participants provide corrective feedback by turning the steering wheel in the direction they desire the car to go for each of the calibration tasks. My approach utilizes the data gathered from the calibration tasks to train the MIND MELD architecture.

My MIND MELD network architecture simultaneously learns the personalized embedding, $w^{(p)}$, describing the way in which an individual is suboptimal (e.g., over- or under-correcting in a driving simulator) and the difference between the optimal corrective

label and the participant provided label. Via variational inference, I maximize a lower bound on mutual information between $d_t^{(p)}$ and $w^{(p)}$, ensuring that $w^{(p)}$ can represent various and distinct feedback styles. Thus, I train $w^{(p)}$ to capture salient information about an individual's style by utilizing the variational lower bound.

Results and Contributions: I first evaluate MIND MELD's ability to improve upon LfD in a synthetic study. In this study, I create artificial calibration tasks, optimal labels, and demonstrator data. I then train the MIND MELD architecture on this data to learn the personalized embeddings and network parameters. Our results show that we are able to improve upon suboptimal labels by 61% on the training tasks and 55% on holdout test tasks. Additionally, I demonstrate that MIND MELD learns meaningful representations of the demonstrator's stylistic tendencies.

I next illustrate MIND MELD's ability to outperform prior work in both human-centric and robot-centric LfD in a human-subjects study. I recruit 76 training participants to complete sixteen calibration tasks which we use to learn the parameters, θ, ϕ, and ϕ' of the MIND MELD network. I then recruit 42 testing participants who complete the calibration tasks and a holdout test task. I demonstrate that MIND MELD is able to achieve a higher probability of reaching the goal throughout the course of the study compared to baseline approaches. Additionally, MIND MELD achieves a lower average distance from the goal.

In this work I demonstrate a personalized framework that enables robots to learn about the way in which an individual demonstrator is suboptimal when teaching a robot. My MIND MELD framework is able to effectively map various suboptimal demonstration styles to better demonstrations. I show that, by adapting to the individual demonstrator, a robot that learns via my MIND MELD algorithm is able to outperform baseline approaches.

1.3.2 Personalized Teaching via Reciprocal Mutual Information Driven Meta-Learning from Demonstration

In my MIND MELD work, I demonstrated that humans provide both suboptimal and heterogeneous demonstrations and that a personalized, data-driven approach can effectively improve upon this suboptimality. However, by correcting for suboptimality under-the-hood, the human demonstrator will never learn how to provide higher-quality demonstrations and MIND MELD will likely only reinforce suboptimal tendencies. This lack of transparency may pose problems when the demonstrator provides demonstrations in a different robotic domain that does not have the ability to correct for suboptimality. Furthermore, as shown in prior work, humans may generalize better to out-of-distribution tasks than a machine learning algorithm and, therefore, it is advantageous for novice human to learn how to become better demonstrators [28].

To solve this problem, I posit that robots should be capable of working with their human demonstrator and providing personalized feedback to help the demonstrator to better understand how to provide high-quality demonstrations. By doing so, human demonstrators will be more likely to succeed in teaching robots in a wide range of tasks across various domains. This reciprocal teaching paradigm allows human demonstrators to effectively teach agents while simultaneously learning how to become better demonstrators.

Approach: To teach humans to become better demonstrators, I expand upon our MIND MELD framework and utilize the personalized embedding learned via MIND MELD as a proxy for suboptimality which I then translate into meaningful and constructive feedback for the demonstrator. Our approach to teaching humans how to become better demonstrators consists of three components: 1) a semantically meaningful personalized embedding space which describes the way in which the individual is suboptimal, 2) robotic feedback to communicate to the demonstrator the way in which the demonstrator is suboptimal, and 3) an ability to update an individual's personalized embedding after receiving robotic feedback.

Semantically Meaningful Embedding Space: I demonstrated in prior work [1, 2] that our

MIND MELD architecture can learn personalized embeddings that correlate with suboptimal stylistic tendencies. In the driving simulator domain, I showed that the embeddings correlate with a participant's tendency to over-/under-correct and provide delayed/anticipatory feedback. I utilize this knowledge about where the demonstrator falls within the embedding space to translate a demonstrator's personalized embedding into robotic feedback.

Robotic Feedback: To determine the feedback that the robot should provide to the human, I calculate the distance along the semantically meaningful dimension between the personalized embedding, $w^{(p)}$ and the optimal embedding, w^*. Because the MIND MELD architecture outputs the difference between the participant's corrective labels and the ground truth, the optimal embedding is defined as the embedding which minimizes the output of the MIND MELD architecture. The robot then provides feedback that is proportional to the distance from the optimal embedding. The robot continually provides feedback until the demonstrator's teaching quality has sufficiently improved.

Updating Embedding: To determine when the demonstrator's embedding has moved sufficiently close to the optimal embedding, we must update the personalized embedding after each iteration of robotic feedback. To do so, we train a Long-Short Term Memory Network (LSTM) architecture to approximate the new embedding based on the demonstrator's recent corrective feedback. This method enables us to dynamically update the robot's understanding of the teacher's suboptimality and adapt feedback accordingly.

Results and Contributions: To evaluate the ability of Reciprocal MIND MELD to improve upon a teacher's ability to provide demonstrations, I first conduct a human subject study in which I show that, given robotic feedback, we can shift a participant's learned embedding towards the optimal embedding in the over-/under-correcting dimension. I evaluate how the quality of demonstrations changes after receiving cooperative, adversarial, and no feedback in a between-subjects study. I find that cooperative feedback shifts a participant's embedding significantly closer to that of the perfect demonstrator, adversarial shifts the embedding significantly farther away and no feedback has little effect on the

embedding.

Next, I conduct a study to investigate how best to provide feedback to improve upon suboptimality in multiple dimensions (i.e., over-/under correcting and anticipatory/delayed). To do so, I conduct a between-subjects study in which participants experience simultaneous feedback (feedback is presented for both dimensions at the same time), greedy feedback (feedback is presented for only the worst dimension), and no feedback. After participants receive multiple rounds of feedback, I find that simultaneous results in significantly greater improvements in demonstration quality.

After demonstrating that Reciprocal MIND MELD can shift a participant's personalized embedding via robotic feedback, I lastly conduct a study showing that we can estimate the demonstrator's new personalized embedding after robotic feedback without having participants redo the calibration tasks. Additionally, I find that providing robotic feedback results in better learning outcomes for an agent and that providing robotic feedback in multiple dimensions simultaneously to the human demonstrator results in the agent achieving a lower average distance from the goal.

In this work, I contribute a personalized teaching framework which learns about the way in which an individual demonstrator is suboptimal and provides personalized feedback to the demonstrator. Our framework is capable of estimating a participant's embedding on the fly and results in improved demonstrations and learning outcomes for an agent.

1.3.3 Manipulating Autonomous Vehicle Embedding Region for Individuals' Comfort

In LfD, end-users tend to be heterogeneous in both their preferences and the way in which they are suboptimal when demonstrating these preferences. In our Reciprocal MIND MELD work, I introduced an approach for personalized teaching to address the problem of suboptimality in LfD. In this next work, I investigate how we can learn about an end-user's preferences and adapt accordingly. One domain in which user preference is an important consideration is autonomous vehicles. Prior work has shown that personalization

of autonomous vehicles (AV) may significantly increase trust, use, and acceptance [29, 30]. In particular, I hypothesize that the similarity of an AV's driving style compared to the end-user's driving style will have a major impact on end-user's willingness to use the AV.

To investigate the impact of driving style on user acceptance, I 1) develop a data-driven approach to personalize driving style and 2) demonstrate that personalization significantly impacts attitudes towards AVs. My approach learns a high-level model that tunes low-level controllers to ensure safe and personalized control of the AV. The key to our approach is learning an informative, personalized embedding that represents a user's driving style. My framework is capable of calibrating the level of aggression so as to optimize driving style based upon driver preference. Across two human subject studies (n = 54), I first demonstrate that our approach mimics the driving styles of end-users and can tune attributes of style (e.g., aggressiveness). Second, I investigate the factors (e.g., trust, personality etc.) that impact homophily, i.e. an individual's preference for a driving style similar to their own.

Approach: To personalize the driving style of an AV, I create a deep learning architecture which simultaneously learns the personalized high-level controller parameters for low level controllers while also learning a personalized embedding describing the driving style of the end-user. I train this network by collecting data from the end-user driving a vehicle. The high-level control parameters predicted by the neural network are fed into low-level controllers which execute lane changes, and maintain velocity and following distance. I utilize the Toyota Research Institute (TRI) driving simulator which is a high-fidelity simulator with a 6-DOF platform.

Prior work has shown that the level of aggression that an AV exhibits has a large effect on preference for the driving style. To test this hypothesis, in this work, I additionally investigate the impact on end-user preference of altering the AV's level of aggression. Do accomplish this objective, I add an additional network head to our personalized network to predict the subjective aggressive style of an end-user. This allows us to shift an end-user's personalized embedding within embedding space to increase or decrease aggressiveness

while holding other factors constant.

Results and Contributions: I conduct a human subjects study in which we observe a participant's driving style and learn a personalized embedding from data collected of their driving. The participant then experiences a Mimic condition in which the AV utilizes their embedding to mimic their own driving style, an Aggressive condition in which the AV exhibits a more aggressive driving style, and lastly a Cautious condition in which the AV exhibits a more cautious driving style.

We find that the Mimic condition generates driving styles consistent with end-user styles ($p < .001$) and participants rate our approach as more similar to their driving style ($p = .002$). Interestingly, we find that a large number of participants preferred a driving style different from their own. We find that personality ($p < .001$), perceived similarity ($p < .001$), and high-velocity driving style ($p = .0031$) significantly modulate an individual's preference for a driving style different from their own.

In this work, I contribute a data-driven approach for personalizing the driving style of an AV. I demonstrate that we are able to successfully mimic the driving style of an end-user as well as produce more aggressive and more cautious behavior. Lastly, I investigate the factors that impact the effect of homophily and find personality, high-velocity driving style, and perceived similarity to be important factors.

1.3.4 Safe Meta Active Learning for Deep Brain Stimulation

In my work on MIND MELD, Reciprocal MIND MELD, and MAVERIC, I introduced approaches to personalize robot behavior and account for both heterogenous suboptimality and preferences of end-users. However, there are many domains which require machines to adapt to end-users while also considering the safety of the end-user. For example, in an AV domain, we must ensure safety of the passenger. I accomplish this goal in our MAVERIC framework by setting safe bounds on the outputs of our neural network (e.g., following distance must be greater than a safe threshold). However, in some domains such

as healthcare, a one-size-fits-all definition of safety may not be the best solution. Instead, machines must be capable of reasoning about the safety of individual patients and adapting safety constraints accordingly.

To address this problem, I introduce a personalized algorithm which explicitly reasons about safety of an individual end-user in a healthcare setting. In this work, I am motivated by the problem of DBS. The domain of DBS exemplifies many of the challenges faced in human-machine interaction; the machine must be capable of understanding and dynamically adapting to the patient to ensure patient health and safety. DBS is a cutting-edge approach to treating otherwise intractable epilepsy that cannot be controlled via pharmacological methods. Surgeons currently employ a manual trial-and-error process to find control settings to reduce seizure frequency. However, there is no standard mapping from parameter values to reduction in seizures that applies to all patients. The optimal stimulation parameter settings can depend on the placement of the device in the brain, the anatomy of an individual's brain, and other confounding factors. Further, a latent subset of parameters can cause negative side-effects. Therefore, there is a need for an approach to efficiently learn the optimal DBS parameter setting for individual patients while also ensuring patient safety.

Approach: I describe our problem set-up in the context of efficiently selecting the optimal parameter setting for DBS in rat models. Rat models are commonly used to gain a better understanding of disease mechanisms and treatments that are unsafe to test in humans [31]. In keeping with [18], I create simulation environments for six rats where, at each DBS parameter setting, the cognitive function of a rat, as described by a "memory score" was measured and dissimulated into a digital twin of the rat. The task is to determine the DBS parameters (e.g., signal amplitude) in the simulation environments that maximize each rat's memory scores (i.e., ability to recall the location of different objects) without causing unwanted side effects (e.g., memory deficits or seizures). Memory deficits and seizures are indicated when the memory score drops below zero.

Safe Meta-Learning Architecture - Our architecture achieves three key objectives: (1)

efficiently learning the optimal parameter setting via active learning, (2) incorporating knowledge from the patient population via meta-learning, and (3) ensuring patient safety through safety constraints imposed via a mixed integer linear program (MILP) that are based on the machine's understanding of the patient model.

Active Learning - To efficiently learn the optimal parameter setting for an individual rat, we need to select the set of DBS parameters that will provide maximal information when learning the model of a rat brain. Our active learning acquisition function describes the amount of information gained when applying a DBS parameter given the parameter-brain state history of the rat. Intuitively, we want to choose the set of DBS parameters that inform us most about the mapping from parameters to memory score, conditioned on the set of parameters and resultant memory scores that the rat has already experienced.

Meta-Learning - Prior approaches [32, 33] have utilized heuristics to quantify information gain. However, these heuristics are only proxies for true information gain and may not accurately quantify the actual information gain of the model when updating the model with new information. Additionally, such heuristics may not be the best metric in a DBS domain. Instead, I utilize meta-learning to learn the acquisition function which describes the expected information gain when selecting a candidate parameter. To learn the acquisition function and encoding of patient history, I meta-learn over a population of rats. Our acquisition function, learns to map a candidate DBS parameter to a measure of information gain (i.e., future expected discounted reward) conditioned on an embedding of sample history. To learn the representation of patient history, I utilize an LSTM neural network, which maps the history to an embedding.

Safety Framework- To ensure safety of the rat, I embed our acquisition function in a chance-constrained linear program. By doing so, I ensure with probability $1 - \epsilon$ that the rat remains within a volume of safety, This volume is parameterized by a safe state (i.e., no seizures) and radius that encompasses all known safe states. This volume can be conservatively converted into linear constraints as an inscribed d-orthotype, thus creating

a convex optimization problem. Our safety objective is to ensure that the probability of remaining in a safe configuration remains higher than $1 - \epsilon$.

Results: I empirically validate that Safe MetAL outperforms baselines in terms of its ability to safely and actively learn the latent parameters of a rat brain model.

Active Learning – Results in our DBS domain empirically validate that our algorithm more efficiently learns the optimal parameters compared to baseline approaches. Figure 6.3 shows the accuracy with which each of the algorithms selects the optimal parameters at each time step, t. Safe MetAL selects a set of parameters that results in $>58\%$ higher information gain compared to our two Bayesian baselines and $>41\%$ higher information gain compared to our active learning baselines. This large increase in information gain that Safe MetAL is able to achieve compared to hand-engineered heuristics, suggests that the meta-learning aspect of Safe MetAL is vital for synthesizing a precise, task-specific acquisition function. Lastly, I show that Safe MetAL outperforms by 47% our meta-learning baseline, LAL, which meta-learns over hand-engineered features. These results demonstrate that our meta-learned embedding is more capable of extracting salient information than the hand-engineered features in LAL.

Safety - Because Safe MetAL is able to more quickly learn the optimal parameter settings, it is also able to ensure safe operation to a greater degree than the baselines. Safe MetAL achieves a 6.3% higher guarantee of safety compared to Maximizing Diversity Schrum2020 for the information gain achieved in Figure 6.3. Safe MetAL achieves 98% greater information gain compared to Epistemic Uncertainty Hastie2017 while achieving an equivalent safety guarantee.

Computation Time - The computation time of active learning algorithms can be of critical importance especially in a healthcare setting in which efficient computation can reduce patient suffering. In the DBS environment, BaO has a slight advantage in computation time, but Safe MetAL trades the time for 58% greater information gain while finding solutions in $\frac{1}{10}^{th}$ of a second. Additionally, Safe MetAL is 68x faster than LAL and 61x faster than

Meta BO, our two meta-learning benchmarks and more than 97% faster than our Bayesian baselines.

In this work, I contributed Safe MetAL, a personalized approach for meta-learning an acquisition function. I proposed a method for embedding this acquisition function in a chance-constrained linear program to impose probabilistic safety guarantees that are based on the learned model of the patient. In a novel DBS domain, I demonstrated the ability of Safe MetAL to safely and efficiently learn the optimal parameter settings for a DBS patient.

CHAPTER 2
BACKGROUND AND RELATED WORK

Personalization of machines to meet the needs of human end-users is necessary in many domains in which humans and machines must interact. Each domain presents unique challenges that the machine must overcome and each domain requires the machine to reason about specific aspects of the end-user so as to optimize the human-machine relationship. For example, in an AV domain, the AV may need to learn about the end-user's driving style and understand the end-user's attitude towards AVs so as to optimize interaction whereas in a healthcare domain, the machine must consider the patient's biology, medical history, and safety.

In this chapter, I focus on prior work in personalization in the technical domains and applications relevant to my thesis work. I first begin with a discussion of previous work in learning from heterogeneous and suboptimal demonstrators and illustrate the need for a personalized framework to fill the gap in prior literature. I then continue by reviewing how prior approaches have attempted to correct for end-user suboptimality via personalized teaching frameworks and demonstrate the need for a personalized framework for providing feedback to human demonstrators in LfD. Next, I investigate the importance of personalization of driving styles of AVs as illustrated in prior work and discuss the lack of prior work on the effect of homophily with regards to personalized driving styles. Lastly, I present prior work in the domain of healthcare in which human-machine systems must also reason about safety in addition to personalization. In this review, I illustrate how my insight of learning from both the individual as well as the population can improve upon the ability of machines to personalize their behavior for specific end-users and thereby improve the human-machine relationship.

2.1 Learning from Demonstration

The objective of LfD is to learn a policy from a set of demonstrations provided by a human teacher [14]. LfD allows non-expert, non-roboticists to teach robots novel tasks without the need for robotic expertise or programming experience [34]. One of the many challenges of creating successful LfD frameworks that can operate in the home and learn from novice users is the heterogeneity of end-users [35]. Differing skills, preferences, and experience with the robot can produce heterogeneity amongst the population of demonstrators. End-users tend to be heterogeneous in an LfD paradigm in two key aspects. Demonstrators tend to be heterogeneous in 1) their preference for how a task should be accomplished and 2) their ability to provide high-quality demonstrations [25]. This heterogeneity violates many assumptions that are held in prior work on LfD. To rectify this problem, prior work has introduced approaches for learning from heterogeneous and suboptimal demonstrators [36, 37, 16].

For robots to successfully learn from humans, they must be capable of taking into account various demonstrator strategies and preferences for how to accomplish a task. For example, when demonstrating how to play table tennis to a robot, one end-user may prefer to demonstrate a top-spin stroke whereas another end-user may demonstrate a side-spin. Because of these differing styles of play, the demonstrations provided to the robot by each end-user will differ and the robot must be capable of reasoning about these differing styles and preferences when learning from diverse users. Prior work has investigated how to account for individual differences in preferences and styles. For example, Chen et al. propose an approach to simultaneously learn the task goal and the human's preferred strategy via network distillation [36]. This approach allows robots to learn from diverse strategies.

While accounting for heterogeneous preferences is an important problem in LfD, in my work, I focus on the problem of heterogeneity with regards to teacher performance. Improving upon an agent's ability to learn from poor quality demonstrations is an important

problem that must be addressed for LfD to be deployed in the real world. In this section I discuss prior work in LfD and specifically focus on prior work in learning from suboptimal demonstrators. I discuss prior work in both inverse reinforcement learning and imitation learning and illustrate why robot-centric LfD demonstrates potential for improving upon a robot's ability to learn from novice human demonstrators. My key insight is that prior work fails to effectively model human heterogeneity in robot-centric LfD which reduces the potential efficacy of robot-centric approaches [38]. In this thesis, I propose a framework to overcome this limitation and demonstrate that a robot-centric LfD approach which accounts for both demonstrator suboptimality as well as heterogeneity is able to outperform prior work in both human-centric and robot-centric LfD.

2.1.1 Inverse Reinforcement Learning for Suboptimal Demonstrators

The goal of Inverse Reinforcement Learning (IRL) is to learn a reward function that best explains the set of human demonstrations and then determine the optimal policy that maximizes the future expected reward [39]. A limitation of many IRL approaches is that they assume perfect demonstrations are provided by an expert demonstrator [40, 39]. Noisy and suboptimal demonstrations can make it difficult to recover the true reward function especially when this suboptimality is not explicitly accounted for. This assumption of optimal demonstrations has been relaxed in more recent approaches [39]. For example, approaches such as T-Rex and D-Rex attempt to improve upon the ability of agents to learn from poor human demonstrations by learning a reward function from a set of ranked demonstrations [37, 16]. Chen et al. introduced SSRR to learn from suboptimal demonstrations by characterizing the relationship between noise and performance [41].

While accounting for suboptimality has improved the ability of IRL approaches to learn from novice and suboptimal demonstrators, IRL approaches still have limited real world applicability. IRL requires that a Markov Decision Process (MDP) be solved which can be computationally costly. Additionally, most approaches require access to the dynamics model.

These requirements can make the deployment of IRL frameworks practically difficult [39].

2.1.2 Imitation Learning

As an alternative to IRL, imitation learning directly learns the mapping from states to actions from human demonstrations via supervised learning. Imitation learning can be framed as a MDP sans reward function (MDP\R). The MDP\R is defined by the 4-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma \rangle$. $\mathcal{S}$ represents the set of states and $\mathcal{A}$ the set of actions. $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S}' \rightarrow [0, 1]$ is the transition function that returns the probability of transitioning to state, s', from state, s, when applying action, a. γ weights the discounting of future rewards. LfD seeks to synthesize a policy, $\pi : \mathcal{S} \rightarrow \mathcal{A}$, mapping states to actions to maximize the future expected reward. In an LfD paradigm, a demonstrator provides a set of trajectories, $\{(s_t, a_t), \forall t \in \{1, 2, ...T\}\}$, from which the agent learns a policy.

Imitation learning can be more feasible to implement in the real-world and does not require access to a transition function to solve an MDP [42]. However, imitation learning approaches make the assumption that the training set and testing set are independent and identically distributed [14]. Two types of imitation learning are human-centric LfD and robot-centric LfD [38]. In human-centric LfD, the human drives the interaction and directly demonstrates the task to the robot. Alternatively, in robot-centric LfD, the agent rolls out its learned policy and the human demonstrator provides corrective feedback [38].

Prior work has explored human-centric LfD for learning a robot policy for task execution from an expert human demonstrator [43, 44, 45, 14, 46]. The simplest and most ubiquitous form of human-centric LfD is Behavioral Cloning (BC), in which a robot infers the mapping from states to actions via supervised learning from human demonstrations. [47, 48]. However, if the learner deviates from the demonstrated path, covariate shift occurs due to a mismatch between the states induced by the demonstrations and those experienced by the robot when rolling out a policy. The mismatch between state distributions violates the i.i.d assumption. Due to this covariate shift, the number of mistakes a learner makes can

compound quadratically in the time horizon [45].

In response to this problem, Ross et al. introduced Dataset Aggregation (DAgger), a robot-centric LfD approach that aggregates a training dataset of expert labels queried during policy rollout [45]. DAgger utilizes the state distribution induced by the current policy to solicit labels from the expert and employs a gating function to determine the mixture of expert and learner during each rollout. Ross et al. proved linear-loss, no-regret guarantees and showed that with high-quality, expert demonstrations, DAgger outperforms prior work [45].

One of the drawbacks to robot-centric LfD is that these approaches require a heavy workload from the demonstrator, which can result in demonstrator fatigue and poor training results [49, 50, 51]. To improve teacher-learner interactions, prior work has attempted to reduce the amount of corrective feedback required of the demonstrator by DAgger [52, 49, 53, 54]. He et al. proposed an imitation-learning-by-coaching algorithm in which the learner must imitate actions of progressively increasing difficulty [52]. In this approach, task loss is reduced by demonstrating to the learner preferable actions. Results have shown that this coaching scheme can outperform DAgger and achieve a lower regret bound when the demonstrator is an oracle, but no study has been conducted demonstrating this method's advantage with human teachers. Rather than requiring demonstrators to provide corrective labels or direct demonstrations, Knox and Stone developed TAMER, which allows humans to provide feedback in the form of a scalar reward [55]. TAMER accounts for delayed feedback, but does not account for heterogeneous demonstrators.

In related work, Kelly et al. proposed to reduce expert workload while improving upon expert-provided demonstrations through Human-Gated Dataset Aggregation (HG-DAgger), allowing the expert to decide when to provide feedback via a gating function [49]. HG-DAgger learns a stationary policy such that labels are obtained via a policy that stabilizes around expert trajectories. Spencer et al. expanded on this idea, utilizing both information about when the expert does and does not intervene, in the Expert Intervention Learning (EIL)

algorithm [53]. HG-DAgger and EIL both focus on augmenting *when* the human should provide feedback during a trajectory, whereas in my work, I focus on *how*, by improving the feedback itself.

While robot-centric LfD approaches such as DAgger and its variants perform well when demonstrations are high-quality, Laskey et al. [38] illustrated that robot-centric learning approaches can lead to human mislabelling, resulting in poor learner performance. Laskey et al demonstrated this short-coming in a human-subjects study in which the authors compared robot-centric and human-centric LfD approaches. In their work, the authors found that, due to the suboptimality of human demonstrators, human-centric and robot-centric LfD perform at parity despite robot-centric LfD's stronger theoretical guarantees. The work by Laskey et al. suggests that one of the main hurdles preventing robot-centric LfD from reaching its full potential is the inability of current approaches to account for suboptimal demonstrations.

2.1.3 Conclusion

Prior work has demonstrated that robot-centric LfD can effectively learn a policy when demonstrations are high-quality, yet humans often struggle to provide good demonstrations [50, 37]. As discussed above, prior work has attempted to account for human suboptimality in LfD [37, 56, 57, 58]. However, there is a lack of prior work that accounts for *heterogeneity* and suboptimality of novice human demonstrators in robot-centric LfD. Reasoning about heterogeneity and suboptimality with regards to demonstrator performance in robot-centric LfD is crucial for the robot to be able to effectively learn from novice demonstrators. Therefore, there is a need for LfD algorithms that can effectively learn from the typical, non-expert human demonstrator in a robot-centric paradigm while taking into account demonstrator heterogeneity [14].

In this thesis, I improve upon a robot's ability to learn from heterogeneous demonstrators in robot-centric LfD. I introduce an approach which captures information about demonstrator heterogeneity and suboptimality via a personalized embedding and utilizes this informa-

tion to improve upon the quality of demonstrations. My approach is the first to improve upon robot-centric learning by inferring demonstrator suboptimal style via personalized embeddings to correct for suboptimal demonstrations. My aim is to maintain the advantages of robot-centric learning (i.e., reducing covariate shift) while making robot-centric LfD more human-aware by accounting for the suboptimality and heterogeneity of human demonstrators. I show that by doing so, we are able to outperform both human-centric and prior robot-centric LfD approaches.

2.2 Personalized Teaching

In the previous section, I surveyed approaches which account for human suboptimality in LfD and discussed why robot-centric LfD approaches need to reason about human heterogeneity and suboptimality to effectively learn from novice end-users. By accounting for heterogeneity, robot-centric LfD will be able to maintain performance guarantees and better learn from suboptimal demonstrators. However, correcting for suoptimality under-the-hood via a machine learning algorithm may not always be the best strategy. In some circumstances it may be beneficial for the machine to provide explicit, personalized instructions to the end-user about how they can improve their performance [59]. Robotic instructions can increase transparency as well as improve the overall perception of the system [59].

Below, I provide evidence from prior work for why instructional feedback may be advantageous in certain circumstances. Additionally, I discuss several studies that have investigated how best to provide feedback to suboptimal demonstrators in an LfD paradigm. This prior work illustrates that providing feedback to suboptimal demonstrators can improve upon their teaching ability. I then survey prior work in developing algorithms for personalized robotic teaching. I show that, despite evidence supporting the need for personalized teaching, there is a lack of prior work in developing personalized robot teaching approaches for suboptimal demonstrators in LfD.

2.2.1 Teaching the Teacher in LfD

LfD enables novice end-users to teach robots novel tasks without the need for programming skills or a robotics backgrounds. However, because of the novice end-user's lack of experience with robots, the end-user may have difficulty understanding how best to provide demonstrations to the robot [60, 59]. The quality of the policy learned by the robot strongly depends on the quality of the demonstrations provided and therefore, robots that receive low-quality demonstrations may struggle to learn. Prior work has assumed that humans may be able to improve their own teaching ability by observing the learned policy of the robot and adjusting their teaching accordingly [34, 61]. Yet, prior work has shown that this may not be effective if the end-user is not able to understand how the performance of the robot's learned policy relates to their teaching. Instead, in many situations, end-users will require additional insight about how to improve upon their demonstrations [59].

Toris et al. conducted a study comparing various LfD methods with novice demonstrators and identified difficulties that arise when non-experts are asked to teach a robot. The authors found that end-users wanted additional insight into what the robot is thinking so as to understand how to provide better demonstrations. The participants specifically stated that they would like the robot to communicate with the them via speech. These findings suggest that simply observing a failed policy may not be enough information for some users to understand how to provide better demonstrations and that communicating informative feedback to the end-user may be beneficial [62].

One option for overcoming the problem of suboptimal demonstrators is to bypass the need for the teacher to improve their suboptimal demonstrations altogether and instead lay the burden on the robot. In such an approach, the robot would be tasked with learning about the way in which the demonstrator is suboptimal and then using this information to learn a better policy. Prior work has investigated a variety of approaches for learning from suboptimal demonstrations [37, 56, 57, 58]. However, in some situations, correcting for suboptimality under-the-hood may not be the best approach. Evidence from the psychology

literature on positive reinforcement suggests that producing a high-quality policy from sub-optimal demonstrations will likely only reinforce the demonstrator's suboptimal tendencies [63]. Furthermore, if the demonstrator is tasked with teaching several robots, not all of which may be capable of correcting for suboptimality, the demonstrator will likely maintain their suboptimal tendencies in each of the robotic domains. Humans also tend to be better at generalizing and extrapolating information to novel tasks than a machine learning algorithm [28].

Thus, I hypothesize that if a human demonstrator is equipped with the skills and knowledge about how to provide high-quality demonstrations then they may be able to produce better results across a variety of tasks and domains than a machine learning algorithm that corrects for suboptimality under-the-hood. There are many factors that must be considered when providing feedback to a human. For example, the domain in which the teaching is taking place, the modality of the feedback, and the content of the feedback must be considered [59]. Additionally, the end-user's familiarity and understanding of how the robot works should be taken into account to ensure that the feedback is provided at the appropriate level of complexity and abstraction.

Several approaches have investigated how to instruct a suboptimal demonstrator on how to improve upon their suboptimal demonstrations. These approaches have demonstrated that actionable feedback can improve a novice demonstrator [60, 59]. Cakmak and Takayama conducted a study investigating several modalities for communicating improvements to a demonstrator. This work compared written instructions with video tutorials in an unstructured setting with naive users. The authors found instructional videos to be the best modality for improving demonstrators' teaching abilities [60]. Sena et al. investigated video feedback and video feedback with rule guidance and found that both modalities produced better results than no feedback [59].

Much of the prior work in this space has investigated tutorials to instruct the end-user before they begin teaching. Providing tutorials and instructions at the beginning of the

teaching process is effective for orienting the end-user on how to best to teach the robot. However, the end-user may require additional instruction while teaching or stray from the best teaching strategy and require correctional instruction. There is little prior work investigating how to provide personalized and adaptive real-time feedback during teaching to adjust to the instructional needs of the human demonstrator.

2.2.2 Personalized Teaching

Outside of the domain of LfD, prior work has produced novel approaches to providing personalized instruction to end-users via robotic tutors [64, 65]. This prior work can provide us with insight into the potential benefits of personalized robotic feedback to improve demonstrator's abilities in LfD. For example, prior work has demonstrated that personalized instructions positively impact learning outcomes for a student tasked with completing a cognitive task [64]. To improve a robot's ability to engage students, Szafir et al. developed an adaptive approach which is capable of measuring students' engagement in real time via electroencephalography and providing personalized verbal and nonverbal cues to regain attention [65] . The authors show that this personalized approach was able to improve recall in students by 43%.

Similarly, Gordan et al. [66] created a personalized method for teaching children a second language via a robotic tutor. Using a reinforcement learning strategy based on valence and engagement cues, the robot was able to learn a personalized strategy to teach the student. The system produced positive results in terms of the number of new words learned and the valence of the children. Leyzberg et al. introduced a personalized system for teaching students via a Hidden Markov Model [67]. This approach was capable of selecting the appropriate lessons for students based on skill and resulted in large improvements in skill. The success of personalized teaching approaches exhibited in prior work suggests that a personalized approach could be beneficial for teaching a novice human demonstrator to be a better robot teacher.

2.2.3 Conclusion

Prior work has demonstrated the importance of personalization when developing autonomous systems capable of teaching end-users. Personalization is particularly important when instructing an end-user in an LfD paradigm, as demonstrators tend to be heterogeneous in the way in which they are suboptimal. Yet, much of the prior work has investigated only one-size-fits-all teaching tutorials that are not capable of adapting to the instructional needs of the individual end-user in real time. In my work, I aim to fill this gap by creating a personalized teaching approach which captures the way in which an individual demonstrator is suboptimal and provides personalized feedback in real-time.

2.3 Autonomous Driving

In the previous sections, I discussed the important considerations for personalization in LfD and surveyed prior work in learning from heterogeneous demonstrators and teaching suboptimal demonstrators how to provide better demonstrations. In this section, I survey work related to a specific application in which personalization is equally important: autonomous driving.

Prior work has shown that human drivers exhibit a vast array of different driving styles. Driving style is defined as the characteristics of driving related to the judgment and decisions of the driver in a specific situation [68]. Prior work has proposed various way to categorize driving style. For example, Taubman-Ben-Ari et al. divide driving style into four different categories: risky, anxious, dissociative, and distress reduction driving [69]. Other work categorizes types of driving styles into aggressive or defensive [70]. Because of these differing driving styles, humans will expect their AVs to drive in a specific manner that is likely related to their own driving style [71]. To meet the expectations of human end-users, AVs must be capable of learning about their end-users and personalizing their driving styles accordingly. Below I discuss prior work in personalization of AV driving styles and how

personalization impacts perception of the AV. Next, I review literature that investigates the question of whether or not AVs should simply mimic an end-user's driving style or if other factors should be considered when determining the optimal driving style for a specific end-user.

2.3.1 Aggressive Driving Style

A common way to categorize driving style is via the level of aggression. Aggression can be measured objectively or subjectively [72, 69]. For example, Bellem et al. quantify driving style via objective metrics including jerk and headway distance [72]. To gain an understanding of a driver's view of their own aggressive style, Harris and Norman developed the Aggressive Driving Behavior Scale [73].

Prior work has investigated the impact of the level of aggression of an AV's driving style on end-user acceptance. For example, Ekman et al. conducted a Wizard-of-Oz study using a Vovlo vehicle with a professional driver demonstrating the various driving styles. The authors compared defensive and aggressive driving styles found that a defensive driving style produced higher trust scores in a Wizard-of-Oz study [30]. Similar work by Basu et al., Yusof et al., and Karlsson et al. found evidence supporting the important impact of the level of aggression on end-user preferences [71, 74, 75]. Because of the large body of literature showing the importance of aggressive driving style, aggressiveness of an AV should be taken into account when designing the optimal AV driving style.

2.3.2 Optimizing Driving Style for Improved End-User Experience

Researchers have demonstrated that personalization of AV driving styles can lead to increased acceptance [30, 29] and may decrease motion sickness [76]. In Sun et al. the authors create personalized driving styles via personalized controllers [29]. The controllers were designed based upon an end-user's own driving style and were intended to mimic an end-user's style. The authors found that this personalized framework increased the sense of

familiarity as well as trust in the system. Ekman et al. investigated the effect of personalization of driving style on trust [30]. The authors compared aggressive and defensive driving styles and measured trust via a mixed methods research design. The authors show that the defensive driving style is considered more trustworthy, in part because participants rate it as more predictable.

Motion sickness is a major concern in autonomous vehicles. Prior work has suggested that motion sickness is caused by a disconnect between expectations and reality during movement [77]. Motion sickness is commonly measured via questionnaires such as the Fast Motion Sickness Scale [78]. Iskander et al. conducted an in-depth review of the causes and consequences of motion sickness in human-controlled and autonomous vehicles [76]. The authors suggest that the way in which the vehicle drives can both exacerbate or decrease the risk of motion sickness. AVs should therefore take into account the end-user's expectations with regards to driving style to reduce the potential for motion sickness.

2.3.3 Personalization Frameworks

To increase trust, likeability and overall acceptance of the AV there is a need for personalized control frameworks which are capable of adapting to the AV driving style of individual end-users. Prior work has introduced a variety of approaches to adapt autonomous vehicles to meet end-user needs and expectations. Many of these approaches aim to mimic an end-user's own driving style. For example, Kuderer et al. utilized an inverse reinforcement learning approach to produce personalized AV behavior via a learned cost function [79]. The authors employ a feature-based reward function learned from human data to mimic the driving style of the end-user. While this method was capable of learning distinct driving styles for different users, the authors did not evaluate their approach in a human subjects study.

Other work investigates personalization of specific aspects of driving [80, 81]. For example, Bolduc et al. developed an approach to match driver's style for adaptive cruise

control [80]. The authors extract parameters from the end-user's own driving and utilize these parameters to inform the cruise control. Feng and Yan explored personalization of lane changes via a support vector machine [81]. By collecting data from lane changes of end-users, the authors train a personalized Support Vector Machine (SVM) to mimic the lane changing style of the end-user. Rather than mimic the driving style of the end-user, Ling et al. introduced a method to adapt driving style online based on the emotional responses of passengers [82]. This approach utilizes EEG signals to analyze the emotions of the passenger and then employs this information to automatically adapt the driving style of the vehicle to match the emotional state. While these approaches have demonstrated that personalizing an end-user's own driving style can lead to increae acceptance and trust in the AV, the question still remains as to whether or not mimicry is the optimal strategy for an AV.

2.3.4 Should We Mimic End-User Driving Styles?

Despite the many approaches that have been developed for mimicking the driving styles of an end-user, prior work suggests that end-users may not want an AV to drive *exactly* as they drive. Instead various latent factors may influence a driver's preference for a specific driving style that may differ from their own driving style [71, 30, 74]. For example, A Wizard-of-Oz study conducted by Yusof et al. found that many end-users prefer a more defensive driving style. In this work, the authors conducted a study using an Audi test vehicle to realistically simulate an AV. Based upon a self-reported questionnaire, participants were categorized as either defensive or assertive drivers. Then each participant experienced a defensive and assertive AV driving style. The authors found that preference for an assertive or defensive driving style depended on the driver's own style and that aggressive drivers prefer more defensive AVs [74].

Basu et al. conducted a study investigating preference for driving styles both similar and different from one's own. The authors investigated a style intended to mimic the participant's driving style, an aggressive driving style, a defensive driving style, and a distractor style.

More than half of the participants preferred a driving style different from their own. The largest predictor for preference was *perceived* similarity suggesting that participants did not want the AV to drive as they drive, but instead preferred the AV to drive like the end-user *thinks* they drive. This finding suggests that, because humans often lack introspection and have a poor perception of their own driving style, their preferred style often differs from their actual driving style [71]. Based on these results, the authors suggest that we can not simply rely on mimicry and instead, must also account for other end-user characteristics to determine the optimal driving style.

2.3.5 Conclusion

Prior work has shown that personalization of AV driving styles is crucial for improving the overall experience for the end-user. The optimal driving style for an end-user is likely related to the end-user's own driving style, and yet it is unclear exactly how it is related. Prior work has suggested that perceived similarity may be important but the impact of other latent factors has not been thoroughly investigated [71]. To answer these questions, we first require a personalized framework that is capable of both mimicking the driving style and modulating style of an end-user based upon relevant subjective factors.

While prior work has investigated approaches for mimicking driving style, no prior work has created an architecture that can modulate driving style with respect to an end-user's own style. Yet, prior work provides evidence that this functionality is important for optimizing driving style for an end-user [71, 74]. Additionally, prior work has not extensively investigated the relationship between subjective factors and preference for styles similar to one's own. In our work, we seek to fill these gaps by proposing an approach capable of producing more or less aggressive behavior with respect to the end-user's own driving style. Additionally, we conduct a thorough investigation into the factors that impact preference for various driving styles.

2.4 Healthcare

Assistive machines, robots, and robotic devices are becoming more common-place in the domain of healthcare to assist clinicians with patient care and make up for a worker shortage. Healthcare applications require these devices to reason about heterogeneous patient biology, differing needs of patients, and the safety of the patient. Due to the diverse biological consideration, disease manifestations, and patient needs, personalization in healthcare is critical to optimize patient care. One crucial component that must be taken into account when developing personalized frameworks in healthcare applications is patient safety. In the previous sections, I discussed personalization approaches in various domains and applications and how these personalized approaches can improve the human-machine system. However, many of these approaches do not explicitly reason about safety of the end-user. In this section I discuss related work in personalization approaches for healthcare and cover prior work that has investigated safe learning in various contexts.

2.4.1 Personalization in Healthcare

Prior work in personalized human-robot interaction has investigated how to personalize patient-machine interaction in a healthcare setting. These approaches require the system to learn about the needs of the patient and adapt accordingly. Polak et al. argued that personalization in the context of rehabilitation robotics is an important goal and that robots must take into account individual preferences as well as the nature of the individual's affliction so as to provide appropriate support and companionship [83].

To address the need for personalization in rehabilitation settings, Tapus et al. introduced a robotic behavior-adaptation system for post-stroke rehabilitation. The robotic system is capable of adapting its interaction to fit the user's personality traits and task performance [84]. In a human subjects experiment, the authors found that robots that matched their behavior and rehabilitation strategy to the personality of the patient were preferred by the

patient. Work by Irfan et al. investigated the benefits of a personalized healthcare robot for cardiac rehabilitation [85]. This approach utilizes sensors to measure patient performance in cardiac related activities and the robot provides personalized feedback based on the sensor data for individual patients. The authors found that personalization increased adherence and motivation. Francois et al. [86] created a framework for robots to adapt to different styles of play via a cascaded information bottleneck method. The authors demonstrated the adaptability of this method in a case study and detail how the method can be applied to therapy for autistic children. This prior work in healthcare indicates that personalization can greatly improve patient care and is important for optimizing patient care.

2.4.2 Safety

When interacting with patients and providing personalized care, autonomous systems must take into account the safety of the patient. Prior work has investigated safe learning in the context of safe Bayesian optimization and safe reinforcement learning. For example, Sui et al. [87] developed the algorithm SafeOpt which balances exploration and exploitation to learn an unknown function; however, this approach has significant limiting assumptions about the underlying nature of the task. Turchetta et al. [88] addressed the problem of safely exploring an MDP by defining an a priori unknown safety constraint updated during exploration, and Zimmwer et al. [89] utilized a Gaussian process for safely learning time series data. However, these approaches do not incorporate knowledge from prior data to increase sample efficiency, limiting their ability to choose the optimal action. Schrum et al. [33] attempted to overcome this problem by employing a novel acquisition function, Maximizing Diversity, which is utilized to quickly learn altered system dynamics in a chance constrained framework. Yet, the hand engineered acquisition function limits the capabilities of this approach.

2.4.3 Active Learning

In healthcare settings, often the machine or robotic system must select the next treatment or diagnostic tool that should be applied to the patient that would provide the most information about the nature of the patient's disease. Active learning is one method by which a system can intelligently and efficiently gather information. My work in personalizing human-machine systems for healthcare (discussed in Chapter 6) utilizes active learning techniques to quickly learn how best to treat the patient. In this section, I review the foundations and prior work in active learning which influence my approach.

Active learning acquisition functions serve as heuristics to select the candidate unlabeled training data sample that, if the label were known, would provide the most information to the model being learned [90, 91, 92, 32]. In Hastie et al. [32], the sample is selected that the learner is least certain about. In work by Ashmaig et al. [18], the authors utilize Expected Improvement (EI) heuristic to balance exploration versus exploitation to determine the optimal stimulation parameters in DBS. Prior literature has also investigated on-the-fly active learning and meta-active learning [93, 94]. Konyushkova et al. [94] describes the algorithm Learning Active Learning (LAL). The authors present a meta-learning method for learning an acquisition function in which a regressor is trained to predict the reduction in model error of candidate samples via hand engineered features. Volpp et. al. [95] alternatively considers a Gaussian Process based method to meta-train an acquisition function on a distribution of tasks. Work by Geifman et al. [96] actively learns the neural network architecture that is most appropriate for a given task, e.g. active learning. Pang et al. [97] additionally proposed a method to learn an acquisition function that generalizes to a variety of classification tasks. Yet, this work has only been demonstrated for classification.

2.4.4 Conclusion

Prior work has demonstrated that personalization in healthcare is important for improving patient care. Additionally, safe learning techniques must be utilized when gathering data

and interacting with the patient to ensure patient safety. Active learning techniques can be employed to efficiently reach a solution. In my work, I take inspiration from prior work in personalization, safe learning, and active learning to develop a personalized framework for efficiently learning the optimal parameters for deep brain stimulation patients. Our work fills the gap in prior work by both personalizating parameter selection while also ensuring safety guarantees.

CHAPTER 3
MUTUAL INFORMATION DRIVEN META-LEARNING FROM DEMONSTRATION

3.1 Introduction

When an individual purchases an in-home cleaning robot, the robot will have to be taught many novel tasks over an extended period of time. The user may have to teach the robot how to move dishes from the dishwasher to the proper location in the cabinets or how to wash the windows and take out the trash. Simply pre-programming these tasks may not be an adequate solution as different users may have differing preferences for how their robot should operate. Therefore, to effectively meet the needs of the end-user, the robot must be capable of successfully learning new tasks quickly via demonstration. Learning from Demonstration (LfD) seeks to enable humans to teach robots new skills via human task demonstrations without the need for users to have prior experience in computer programming [43]. In LfD, the robot learns a policy that maps the state of the world to how the robot should act to accomplish the human-specified or demonstrated task [98]. Researchers have pursued two principle types of LfD: human-centric and robot-centric [38]. In human-centric LfD, a human typically performs the task, and the robot infers from this demonstration the task specification. An example of human-centric LfD is BC, i.e. *mimicry* [99], where the robot records the human demonstration of the task and uses supervised learning to learn a policy mapping states to actions. However, Behavioral Cloning (BC) suffers from covariate shift issues due to a mismatch between the distribution of states given by the demonstration versus those experienced by the robot when attempting to accomplish the task [47, 45, 100].

Robot-centric LfD is an alternative to human-centric LfD and addresses the problem of covariate shift [45] by instead learning from a human's corrective feedback signal at

each time step as the robot executes the task [38]. One example of robot-centric LfD is Dataset Aggregation (DAgger) [45]. Ross et al. showed that learning from human corrective actions solves the problem posed by covariate shift [45]. Many robot-centric, as well as human-centric, LfD algorithms assume the demonstrator is an expert at the task and that they will provide optimal demonstrations or corrective labels [14]. When the demonstrator is a Wizard-of-Oz oracle [26] and provides optimal demonstrations, prior work has shown that DAgger can learn policies that are more sample efficient and accurate than human-centric LfD algorithms [45]. However, these studies may not translate to real-world settings where non-oracle, heterogeneous human demonstrators provide sub-optimal demonstrations [101, 53, 38, 102]. Prior work has shown that humans struggle to provide high quality corrective actions during robot-centric [59]. Additionally, humans are heterogeneous: the way humans provide demonstrations may differ depending upon the individual's abilities and prior experience [103, 104]. Therefore, robot-centric LfD approaches need to account for the teacher's suboptimality and heterogeneity to learn effective policies. However, prior work fails to take into account demonstrator suboptimality and human heterogeneity in robot-centric LfD. To effectively account for the heterogeneity amongst demonstrators, personalized LfD techniques are required to ensure that robotic systems can effectively learn from demonstrators who tend to differ in the way in which they are suboptimal.

To fill this gap, in this chapter, we aim to harness the potential advantages of robot-centric algorithms (i.e., increased policy performance and sample efficiency) and improve upon robot-centric algorithms by explicitly learning to account for heterogeneity and suboptimality in teaching. We introduce Mutual Information Driven Meta-Learning from Demonstration (MIND MELD), which uses an Long-Short Term Memory Network (LSTM) neural network-based architecture to meta-learn a person-specific mapping from human-provided, corrective-action labels to idealized labels, which are inferred based upon a distribution of calibration tasks with known, optimal labels. Because human feedback is heterogeneous, we propose to use variational inference to learn a personalized embedding that encapsulates

information about a person's style of providing corrective demonstrations. We then use the personalized embedding to map each individual's suboptimal labels to labels that more closely approximate optimal labels, thereby improving the performance of robot-centric LfD algorithms. Optimal labels (i.e., ground truths) are only necessary for a small set of calibration tasks [27, 105] to learn to improve upon human labels and are not needed at test time.

In this chapter, we conduct an Institutional Review Board (IRB)-approved within-subjects study, comparing the performance of MIND MELD to a robot-centric baseline, DAgger, and a human-centric baseline, BC. We evaluate these algorithms based on their ability to learn the task of driving an autonomous vehicle to a goal without collisions as well as various subjective metrics. Additionally, we analyze how the learned personalized embeddings capture the demonstrator's style and improve suboptimal labels. Our approach is the first to improve upon robot-centric learning by inferring demonstrator style via personalized embeddings to correct for suboptimal demonstrations. We maintain the advantages of robot-centric learning (i.e., reducing covariate shift) while making robot-centric LfD more human-aware by accounting for the suboptimality and heterogeneity of human demonstrators.

In our work, we contribute the following:

1. We formulate MIND MELD, a novel, personalized LfD framework for improving upon suboptimal corrective labels by inferring individual demonstrator styles.

2. We demonstrate that MIND MELD objectively outperforms prior work in a human-subjects experiment in its ability to reach the goal more often than BC ($p < .001$) and DAgger ($p < .001$).

3. We show that users prefer MIND MELD over DAgger and BC in terms of trust ($p < .001$), workload ($p = .005$), perceived intelligence ($p = .008$), and likeability ($p = .004$).

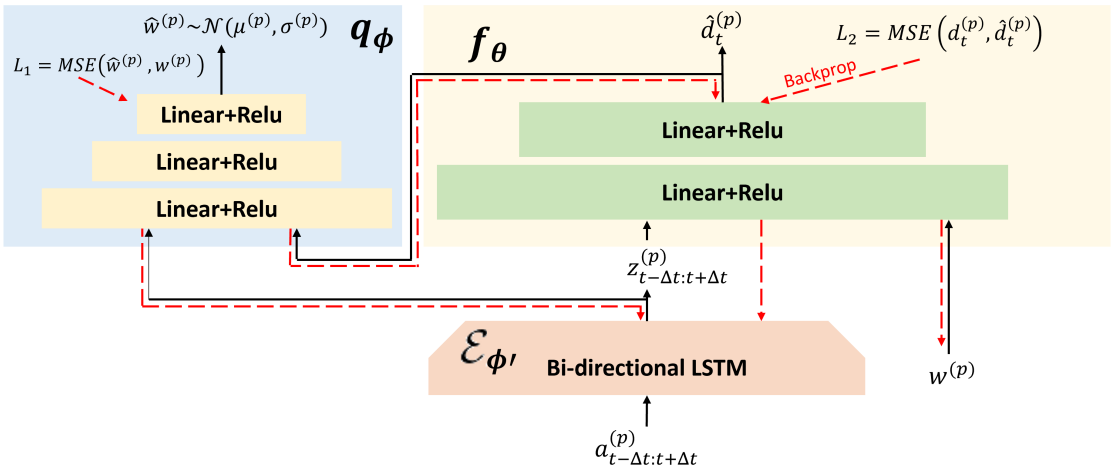


Figure 3.1: This figure shows the MIND MELD network architecture. $a_t^{(p)}$ represents demonstrator p's corrective label, at time t. The recreation subnetwork, q_ϕ, maximizes mutual information between the learned embedding, $w^{(p)}$, the encoding, $z^{(p)}_{(t-\Delta t:t+\Delta t)}$, and the output, $\hat{d}_t^{(p)}$. The objective is to minimize the Mean-squared error (MSE) between the predicted difference, $\hat{d}_t^{(p)}$, and the true difference, $d_t^{(p)} = a_t^{(p)} - o_t$, of the demonstrator's corrective labels and the ground truth label, o_t. We pass in the sequence of corrective demonstrations, $a^{(p)}_{(t-\Delta t:t+\Delta t)}$, from time $t - \Delta t$ to $t + \Delta t$ to the bi-directional LSTM and extract sequential information to inform the predictions of ground truth label at time t.

3.2 Methodology

In the following section, we provide an overview of the preliminaries of our work and describe our MIND MELD algorithm for improving robot-centric LfD with suboptimal human demonstrators. We discuss our network architecture, personalized embeddings, and the mapping of suboptimal labels to more effective labels.

3.2.1 Preliminaries

The LfD problem can readily be framed as a MDP sans reward function (MDP\R). The MDP\R is defined by the 4-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma \rangle$. $\mathcal{S}$ represents the set of states and $\mathcal{A}$ the set of actions. $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S}' \rightarrow [0, 1]$ is the transition function that returns the probability of transitioning to state, s', from state, s, applying action, a. γ weights the discounting of future rewards. LfD seeks to synthesize a policy, $\pi : \mathcal{S} \rightarrow \mathcal{A}$, mapping states to actions to maximize the future expected reward. In an LfD paradigm, a demonstrator provides a set of

trajectories, $\{(s_t, a_t), \forall t \in \{1, 2, ...T\}\}$, from which the agent learns a policy.

We make the following assumptions in our work.

- In the context of robot-centric learning from demonstration, humans provide corrective feedback that is suboptimal (e.g., with respect to an optimal, minimum-jerk, collision-free trajectory planner).
- These human-specified, heterogeneous, sub-optimal strategies can be represented by a learned embedding.
- Across different tasks, humans provide predictable and consistent, albeit suboptimal, corrective demonstrations.
- We have access to a distribution of calibration tasks from which we can obtain the optimal, ground truth labels.

Given these assumptions, we learn an individual's " suboptimal style" via a personalized embedding trained over a set of calibration tasks to represent the human's style. We then utilize this embedding to condition a meta-learned mapping from suboptimal corrective labels to ground truth labels given a set of calibration tasks. Our approach is a type of meta-learning as we learn an architecture over a distribution of tasks and participants in order to more effectively learn a specific LfD task.

3.2.2 Architecture

Depicted in Figure B.3 is the architecture of our network, which consists of three components: 1) the bidirectional LSTM encoder, $\mathcal{E}_{\phi'}: A \rightarrow Z$, 2) the prediction subnetwork, $f_\theta: Z \times W \rightarrow \mathbb{R}$, and 3) the mutual information subnetwork, $q_\phi: Z \times \mathbb{R} \rightarrow \mathcal{N}_W$. The label we aim to improve upon is $a_t^{(p)}$. We denote the set of d-dimensional, personalized embeddings as W, and the set of k-dimensional encodings extracted from the sequences of corrective demonstrations as $Z \subset \mathbb{R}^k$. $\mathcal{E}_{\phi'}$ is trained to extract the encoding, $z^{(p)}_{(t-\Delta t:t+\Delta t)} \in Z$, for the

sequence of corrective labels, $a^{(p)}_{(t-\Delta t:t+\Delta t)}$, provided by person p from time $t-\Delta t$ to $t+\Delta t$.

f_θ maps the encoding, $z^{(p)}_{(t-\Delta t:t+\Delta t)}$, and personalized embedding, $w^{(p)}$, to the difference, $d^{(p)}_t = o_t - a^{(p)}_t$, between the ground truth label (obtained via a controller such as Model Predictive Control (MPC) [106] or Stanley [107]) and the individual's corrective label, where $d^{(p)}_t \in \mathbb{R}^k$. The subnetwork q_ϕ learns a mapping of the encoding, $z^{(p)}_{(t-\Delta t:t+\Delta t)}$, and predicted difference, $\hat{d}^{(p)}_t$, to a posterior distribution over the demonstrator's embedding, $w^{(p)}$. We initialize $w^{(p)}$ based upon the prior, $\hat{w}^{(p)} \sim \mathcal{N}(0,1)$, and obtain an estimate of the individual's learned embedding, $\hat{w}^{(p)}$, by sampling from the approximate posterior.

3.2.3 Variational Inference

This work is motivated by the assumption that humans are not optimal or homogeneous in how they provide feedback, thus necessitating democratized LfD methods which account for both heterogeneity and suboptimality. Note that we handle the fact that individuals' demonstrations are suboptimal and heterogeneous separately. We capture information about an individual's corrective "style" (i.e., *how* they are suboptimal) using a personalized embedding, $w^{(p)}$, for individual p, which we then use to correct the individual's suboptimal and heterogeneous demonstrations, as described in Equation B.1. In our work, we seek to maximize the mutual information between the corrective mapping, $\hat{d}^{(p)}_t$, our learned personalized embedding, $w^{(p)}$, and the encoding of the demonstrator labels, $z^{(p)}_{(t-\Delta t:t+\Delta t)}$, such that the uncertainty of our learned embedding decreases, given informative corrective feedback.

Maximizing mutual information necessitates access to an intractable posterior distribution, $P[w^{(p)}|z^{(p)}_{(t-\Delta t:t+\Delta t)}, \hat{d}^{(p)}_t]$. Thus, we train $w^{(p)}$ to capture salient information about an individual's style by utilizing the variational lower bound, $L_I(f_\theta, q_\phi)$, as derived in Chen et al. [108] and shown in Equation B.1, where the mutual information between $z^{(p)}_{(t-\Delta t:t+\Delta t)}$, $\hat{d}^{(p)}_t$ and personalized embedding, $w^{(p)}$, is $I(w^{(p)}; z^{(p)}_{(t-\Delta t:t+\Delta t)}, \hat{d}^{(p)}_t)$.

$$I(w^{(p)}; z^{(p)}_{(t-\Delta t:t+\Delta t)}, \hat{d}^{(p)}_t) = H(w^{(p)}) - H(w^{(p)}|z^{(p)}_{(t-\Delta t:t+\Delta t)}, \hat{d}^{(p)}_t)$$
$$\geq \mathbb{E}[log(q_\phi(w^{(p)}|z_{(t-\Delta t:t+\Delta t)}, \hat{d}^{(p)}_t))] + H(w^{(p)}) = L_I(f_{\theta,q_\phi}) \tag{3.1}$$

Our network is trained by combining two loss functions: one to learn the embedding, $w^{(p)}$, and one to learn the difference, $\hat{d}^{(p)}_t$, between the demonstrator's corrective label and the optimal label as shown in Figure B.3. L_1 minimizes the MSE between the sampled embedding approximation, $\hat{w}^{(p)}$, and the personalized embedding, $w^{(p)}$ (equivalent to maximizing the log-likelihood of the posterior). L_2 minimizes the MSE between the predicted difference, $\hat{d}^{(p)}_t$, and the true difference, $d^{(p)}_t = o_t - a^{(p)}_t$. We backpropagate the sum of these losses (Equation B.2) to learn the embedding during training such that the personalized embedding reflects the individual's feedback style. Then, at test time, we freeze the network parameters, θ, ϕ, and ϕ' and utilize this personalized embedding to inform the mapping of demonstrator feedback.

$$L_{(\theta,\phi,\phi',w)} = L_{1_{(\theta,\phi,\phi')}} + \lambda L_{2_{(\theta,\phi')}} \tag{3.2}$$

$$L_{1_{(\theta,\phi,\phi')}} = \frac{1}{K+1}\sum_{k=0}^{K} ||\hat{w}^{(p)}_k - w^{(p)}_k|| \tag{3.3}$$

$$L_{2_{(\theta,\phi')}} = ||d^{(p)}_k - \hat{d}^{(p)}_k|| \tag{3.4}$$

3.3 Synthetic Experiment and Pilot Study

We conduct a synthetic study as shown in Figure 3.2 [109] to fine-tune our architecture and demonstrate its efficacy. To do so, we create a set of artificial DAgger-like roll-outs. The ground truth labels are calculated as the difference in heading of the agent and the angle to the goal. We create a set of artificial, suboptimal demonstrators by randomly assigning each

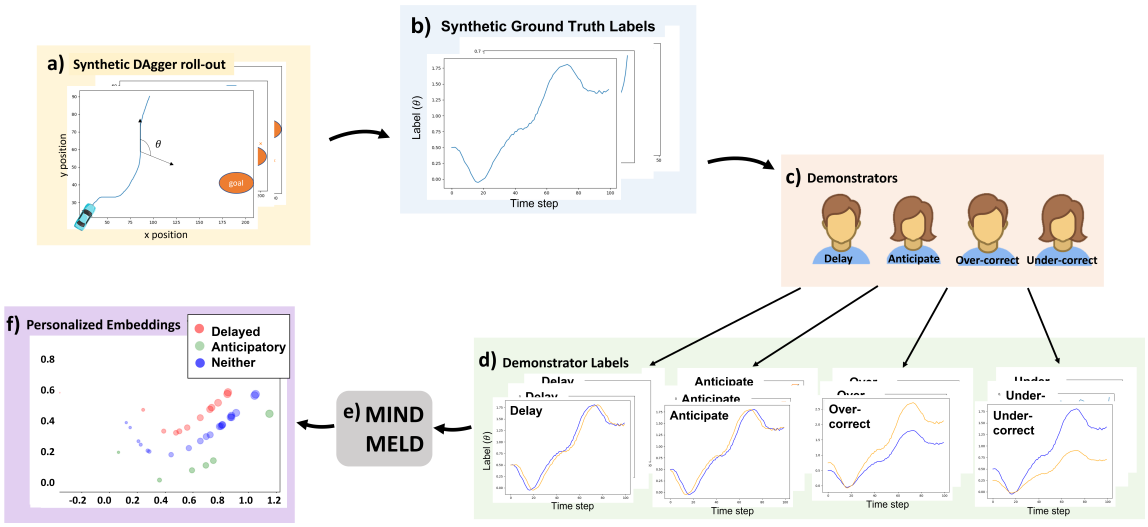


Figure 3.2: This figure shows the creation of the synthetic data. a) shows the artificial DAgger rollouts, b) the ground truth labels, c) the demonstrators, and d) the corrective feedback. g) shows the mapping of suboptimal labels via our architecture,MIND MELD, producing embeddings shown in f). In f), the size of a point represents the degree to which an individual over- or under-corrects. The color represents the individual's style (i.e., delayed, anticipatory, or neither).

demonstrator either a delayed, anticipatory, or neither style and to be either an over-corrector or under-corrector by a randomly selected magnitude. This "style" is then utilized to map the ground truth labels to suboptimal, artificial human labels. We employ this artificial data to demonstrate the ability of our architecture to correct for poor human labels.

3.3.1 Results

Figure 3.2f shows the learned embeddings plotted in latent space. Figure 3.2f shows that the embeddings for individuals that greatly over-correct are clustered towards the right of the graph and those that greatly under-correct are located towards the left. Those who neither over-correct nor under-correct are located at the elbow in the plot. Additionally, those who provided delayed feedback are located towards the top of the plot and those who provided anticipatory feedback are located towards the bottom. Because demonstrators that are similar in the way in which they are suboptimal tend to cluster together, these results confirm that our embeddings learn meaningful representations of an individual's feedback

Algorithm 1 MIND MELD Procedure

For M training participants, collect calibration task data
Perform gradient descent on $\theta, \phi, \phi', \omega$ until convergence (Equation B.2)
Freeze architecture parameters, ϕ, ϕ' and θ
for p in test participants **do**
    Initialize $w^{(p)} \leftarrow \frac{1}{M} \sum_{i=0}^{M} w^{(i)}$
    Collect calibration task data from p
    Perform gradient descent on ω until convergence (Equation B.4)
    Obtain initial demonstration from p.
    Present LfD algorithm conditions {MIND MELD, BC, and DAgger} in randomized order.
    for c in conditions **do**
        Train learner via condition, c, for N demonstrations.
    end for
end for

style. Furthermore, we confirm that our architecture successfully maps the suboptimal labels to labels that are closer to the ground truth embeddings. We find a 61% improvement of labels in the calibration tasks. For unseen test tasks that are not used to train our network, we find a 55% improvement in the quality of the labels after mapping.

We additionally conducted an IRB approved pilot study [109] to test MIND MELD's ability to learn meaningful embeddings and improve upon suboptimal corrective feedback. After recruiting 34 participants, we found that MIND MELD was able to improve corrective feedback and learn embeddings that significantly correlate with demonstrators' stylistic tendencies, i.e., the way in which they deviate from optimal ($p < .001$). Based on results of our pilot study, we redesigned our study to better capture the stylistic tendencies of demonstrators and expanded upon our participant pool.

3.4 Human-Subjects Experiment

We next evaluate our architecture via a human-subjects experiment with human demonstrators and compare our approach to baselines in human- and robot-centric LfD. Through this experiment, we demonstrate MIND MELD's ability to outperform prior LfD work by improving upon a user's suboptimal corrective labels. Our human-subjects experiment

Figure 3.3: The simulator and steering wheel in our human-subjects experiment are on the left and the test task is on the right.

consists of a training phase and a testing phase as discussed below. The steps comprising our study are illustrated in Algorithm Figure 3. Our study has been approved by Georgia Tech's IRB.

Calibration Phase - In the calibration phase, we recruit participants to complete a set of calibration tasks to meta-learn the MIND MELD parameters, θ, ϕ, and ϕ' and personalized embeddings, $w^{(p)}$. Additionally, participants in this phase complete the pre-study questionnaires to capture prior experience and other demographic information.

Testing Phase - For the testing phase, we recruit a set of testing participants for a within-subjects study. These participants first complete the calibration tasks to learn their personalized embedding via Equation B.2. The participants then train an LfD agent via the three learning algorithms, MIND MELD, BC, and DAgger, the order of which is randomized and counterbalanced to mitigate confounding factors (e.g., fatigue, learning effects, etc.). The test task differs from the calibration tasks but is similar and falls within the same distribution (depictions of the calibrations tasks are in the Appendix). The participants in the testing phase complete both the pre-study and post-study questionnaires.

3.4.1 Driving Simulator Domain

We evaluate our approach with a human-subjects experiment in a virtual driving environment, a common domain in prior LfD, HRI, and robotics research [50, 45]. We choose to use the AirSim [110] driving simulator, an Unreal Engine-based high-fidelity physics simulator. Individuals in this experiment interact with the virtual driving environment using an Xbox steering wheel, shown in Figure 3.3. We use a geometric Unreal environment where the LfD objective is to teach the agent to drive to a large, orange ball while avoiding all obstacles. The learning algorithms do not have access to the location of obstacles or the orange ball. We constrain the action space to be the position of the wheel, ranging from -540 degrees to 540 degrees. We define the state space to be composed of an image captured by a camera positioned at the front of the car as well as the car's acceleration, velocity, and position.

3.4.2 Calibration Tasks and Ground Truths

We create a series of sixteen Wizard-of-Oz [26] rollouts which are representative of successful and unsuccessful trajectories and allow us to capture the feedback styles of participants. All participants complete these tasks so MIND MELD can infer their personalized embeddings, w.

To determine ground truth optimal states for each point along the trajectories of the calibration tasks, we employ RRT* [111] (see Appendix for an example). We then apply an MPC controller along the path to determine the ground truth label at each time step.

3.4.3 Conditions

The participants first complete a set of calibration tasks which are used to learn their personalized embeddings for MIND MELD. Then, participants provide an initial demonstration from which all three agents learn an initial policy, π_0. All agents are trained for N demonstrations. Each participant experiences the following conditions in a random order.

Supervised BC - Participants in this condition teach the agent via BC. To mirror our

other conditions, the agent's policy is rolled out with each iteration of training so that the participant can observe the agent's behavior before providing the next demonstration.

DAgger - Participants in this condition teach the agent via vanilla DAgger [45] implemented based on prior work [112, 38]. The agent rolls out policy, π_n, and participants provide corrective labels. The corrective labels are aggregated with the initial demonstration and corrective labels from trials 1 to $n-1$ and the agent is retrained to yield policy, π_{n+1}.

MIND MELD (Ours) - For each demonstration, n, participants provide corrective labels to the agent. This corrective labels are mapped to predicted ground truth labels via MIND MELD. The mapped labels are aggregated with the initial demonstration and mapped labels from trials 1 to $n-1$ and the agent is retrained to yield policy, π_{n+1}.

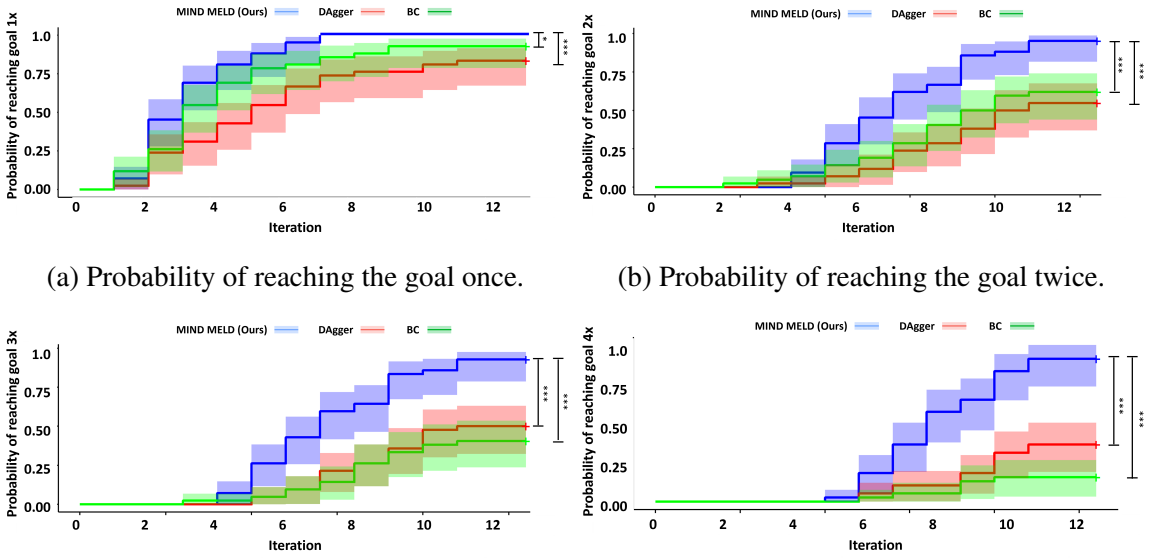
3.4.4 Metrics

Below we discuss the metrics by which we evaluate MIND MELD and the learned embeddings. Both training and testing participants complete the pre-study questionnaires to determine if demographic information correlates with the learned embeddings. Only testing participants complete the post-study questionnaires. The surveys detailed below comply with the design guidelines outlined in Schrum et al. [113] and are validated from prior work when possible. The full text of the surveys and additional surveys that are not relevant to our results can be found in the Appendix. We report Cronbach's alpha (α) for each scale.

Objective Metrics

Stylistic tendencies - We analyzed participants' suboptimality by calculating their stylistic tendencies via Dynamic Time Warping (DTW) [114] between the participant labels, a, and ground truths, o, along two-dimensions: 1) over-/under-correcting (i.e., turning the wheel too far or not enough) and 2) providing delayed/anticipatory feedback. Additional details on our calculations can be found in [109].

Goal Consistency - We measure the total number of times the agent reaches the goal, the number of demonstrations required for the agent to reach the goal, and the probability of



(a) Probability of reaching the goal once.

(b) Probability of reaching the goal twice.

(c) Probability of reaching the goal three times.

(d) Probability of reaching the goal four times.

Figure 3.4: This figure shows that MIND MELD has a statistically significantly higher probability of reaching the goal once (Figure 3.4(a)), twice (Figure 3.4(b)), three times (Fig Figure 3.4(c)), and four times (Fig Figure 3.4(d)) throughout the duration of the study compared to the baselines.

each agent reaching the goal after each demonstration.

Distance - For each policy rollout of the agent, we measure the final distance between the agent and the goal.

Pre-Study Questionnaires

Prior Experience - We collect information about a participant's familiarity and experience playing video games (Cronbach's $\alpha = .93$) and driving a physical car ($\alpha = .93$) via two Likert scales to determine if prior experience correlates with the learned embeddings. Each Likert scale has eight items and a 5-point response format (strongly disagree to strongly agree). Since this survey on prior experience is ad hoc, the Appendix includes a factor analysis to validate the scales.

Post-Study Questionnaires

Trust ($\alpha = .96$) - We measure the participant's trust of the agent after each trial and for each condition [115]. In our results, we analyze the final trust survey from each condition due to the statistical testing considerations detailed in the Appendix.

Workload - We measure the workload after each condition via the NASA Task Load Index (NASA TLX) [116].

Likeability ($\alpha = .95$) - We measure likeability after each condition via the Godspeed likeability subscale [117].

Intelligence ($\alpha = .95$) - We also measure the perceived intelligence of the agent after each condition via the intelligence subscale of Godspeed [117].

3.4.5 Procedure

An overview of our procedure for learning the MIND MELD architecture and validating MIND MELD's ability to outperform our baselines is detailed in Alg. Figure 3. We first recruit 76 training participants by word of mouth and mailing lists. The training participants provide corrective labels for each pre-recorded rollout which we then use to train MIND MELD and learn the parameters of MIND MELD's three subnetworks, θ, ϕ, and ϕ' as well as learn the personalized embedding, $w^{(p)}$, via Equation B.2-Equation B.4. All training participants additionally answer the pre-study questionnaires.

We then recruited 42 different testing participants who experience each of the conditions discussed in Subsection 5.3.4. To learn their personalized embeddings, all participants complete the calibration tasks. We then present each of the conditions discussed in Subsection 5.3.4 in a randomized order. All testing participants complete the pre- and post-study questionnaires.

To ensure that participants are familiar with the system before providing corrective labels, all participants drive around in the simulator for several minutes. Additionally, participants practice providing corrective labels in the first four calibration tasks which are not used in the training of MIND MELD so as to reduce novelty effects.

3.4.6 Hypotheses

Hypothesis 1 - *MIND MELD will improve the corrective labels provided by the participants in the calibration tasks*. We hypothesize that MIND MELD will learn to map suboptimal labels to labels that more closely approximate optimal labels by learning an embedding of stylistic tendencies of individuals.

Hypothesis 2 - *The learned embeddings will correlate with participants' stylistic tendencies and prior experience.* Based on our pilot study [109] illustrating that the learned embeddings correlated with stylistic tendencies, we predict that we will be able to reproduce these results with a larger participant pool. We also predict that the embeddings will correlate with participants' experience with video games and driving.

Hypothesis 3 - *MIND MELD will outperform DAgger and BC in terms of ability to reach the goal.* We hypothesize that, due to MIND MELD's ability to correct for suboptimal labels, MIND MELD will be more likely to reach the goal and achieve a shorter average distance from the goal.

Hypothesis 4 - *The amount by which a participant deviates from the optimal feedback style will correlate with MIND MELD's ability to outperform DAgger.* We hypothesize that participants who provide feedback that differs the most from optimal (i.e., greatly over-correct) will produce poor results for DAgger. Because MIND MELD can correct for this suboptimality, the advantage of our MIND MELD algorithm over DAgger will increase with increasingly suboptimal demonstrations.

Hypothesis 5 - *We hypothesize that MIND MELD will achieve higher ratings on our subjective metrics compared to baselines.* Because MIND MELD corrects for suboptimality, we hypothesize that MIND MELD will be rated higher in terms of perceived intelligence, likeability, workload, and trust.

3.5 Results

We recruited 76 training participants ($M = 22.8$; $SD = 5.5$; 31.2% Female), each of whom completed the calibration tasks and filled out the pre-study questionnaires. We then recruited 42 testing participants ($M = 22.1$; $SD = 2.72$; 40% Female), each of whom completed the calibration tasks, all questionnaires, and experienced the three conditions. In our following analysis, we first determine if the data complies with parametric test assumptions before employing a parametric test. Additionally, we test each model for ordering effects and confounding factors from our covariates and find none. Specific details for all parametric testing assumptions and covariates can be found in the Appendix.

We first test if our findings support **Hypothesis 1** which predicts that MIND MELD will improve upon the corrective labels provided in the calibration tasks. We find a 55% improvement in the labels for our training participants and 37.6% improvement for our testing participants. In the Appendix, we provide graphical depictions of MIND MELD's ability to correct for suboptimal trajectories.

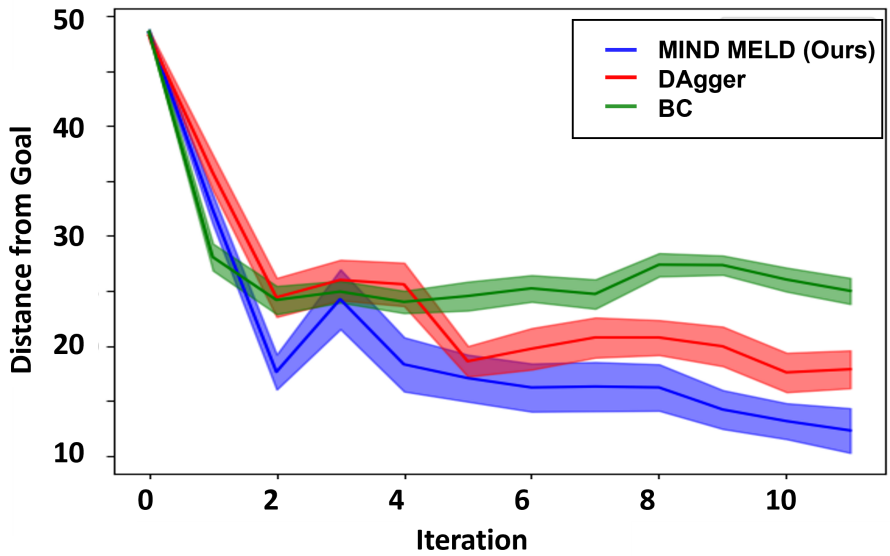


Figure 3.5: This figure shows the average distance and standard deviation from the goal for each algorithm after each iteration. At each iteration, the agent rolls out the current policy and the participant provides a demonstration.

To test **Hypothesis 2**, we conduct a correlation analysis between the learned embeddings and the results of our dynamic time warping describing participants' over-/under-correcting

	MIND MELD-DAgger	MIND MELD-BC	DAgger-BC
Workload	-8.1 (2.8) $p = .005$	-10.0 (2.8) $p < .001$	-2.0 (2.9) $p = .87$
Likeability	1.1 (.25) $p = .004$	1.4 (.28) $p = .001$	.31 (.27) $p = .37$
Intelligence	1.2 (.32) $p = .008$	1.7 (.31) $p < .001$	.53 (.31) $p = .35$
Trust	0.80 (.16) $p < .001$	1.1 (.14) $p < .001$	0.32 (.14) $p = .192$
Distance	-4.5 (.88) $p < .001$	-7.7 (.80) $p < .001$	-3.2 (.82) $p = .01$

Table 3.1: We report the means (standard deviations) of the difference between the agents and associated p-values for objective and subjective metrics.

and delayed/anticipatory tendencies. We find support for the results in our pilot study and find that the learned embeddings significantly correlate with participants' tendency to over-/under-correct ($r(116) = -.47, p < .001$) and provide anticipatory/delayed demonstrations ($r(116) = .49, p < .001$). To further investigate **Hypothesis 2** and determine if prior experience correlates with the learned embeddings, we conduct a correlation analysis between experience with driving and experience with video games. We find that experience with video games significantly correlates with the learned embedding ($\rho = .19, p = .038$).

To investigate **Hypothesis 3**, we next analyze the ability of each agent to reach the goal, in terms of both probability and frequency, over the course of the study. To determine the probability of reaching the goal at each iteration, we conduct a survival analysis, a statistical technique commonly used in medical research to assess the expected time until an event takes place [118]. Survival analysis allows us to analyze data for which an event may never occur. For example, an agent may never reach the orange ball during the study, yet we can still include this data in our survival analysis as "censored" data. In our study, time corresponds to the number of demonstrations that the agent has experienced. An event occurs when the agent reaches the goal the specified number of times.

Figure 3.4 shows the Kaplan-Meier curves for reaching the goal once, twice, three times, and four times. We find that MIND MELD is statistically significantly more likely to reach

the goal once (log rank $p < .001$), twice (log rank $p < .001$), three times (log rank $p < .001$), and four times (log rank $p < .001$) throughout the course of the study compared to DAgger and BC. We find that MIND MELD has a 100% chance of reaching the goal once after the seventh iteration whereas the baselines never achieve 100% probability of reaching the goal even once. Likewise, we find that MIND MELD has a $> 80\%$ chance of reaching the goal three times after the ninth iteration whereas the baselines have a $< 50\%$ chance. This result supports **Hypothesis 3** and shows MIND MELD learns a better policy in terms of probability of reaching the goal.

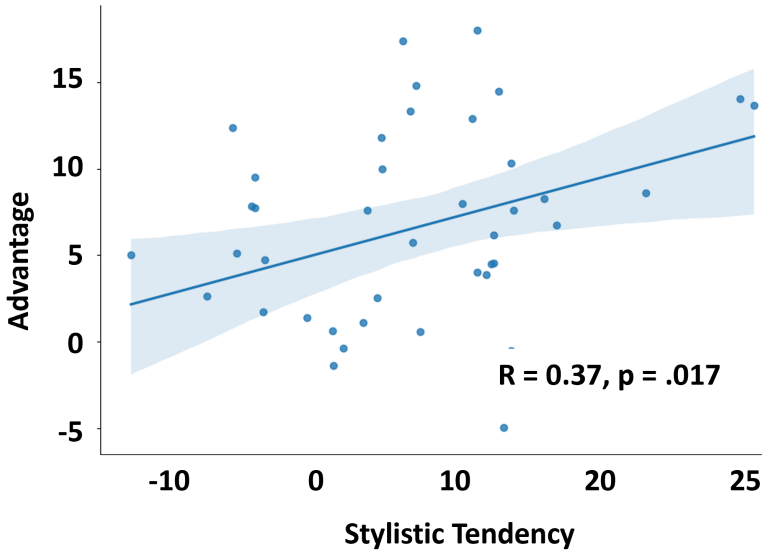


Figure 3.6: This figure shows a plot of participants' tendency to provide delayed/anticipatory feedback vs. the difference between the average performance of MIND MELD and DAgger.

We additionally apply a Poisson regression with a Tukey post hoc to determine if there is a statistically significant difference between the total number of times that each agent reaches the goal throughout the study. We find that MIND MELD reached the goal 2.1x more than DAgger ($p < .001$) and 2.6x more than BC ($p < .001$).

Next, we analyze the average distance from the goal across iterations for each algorithm. We conduct a repeated measures Analysis of Variance (ANOVA) with a Tukey post hoc comparing the distance to the goal for each condition. As shown in Table B.2, we find that MIND MELD achieved a statistically significantly lower average distance from the goal

$(M = 20.4, SD = 5.58)$ compared to DAgger $(M = 24.8, SD = 5.92, p < .001)$ and BC $(M = 28.2, SD = 4.86, p < .001)$. Figure 3.5 shows the average distance to the goal for each trial and condition. Note that a trial ends after the agent either reaches the orange ball or crashes into an obstacle.

To determine if our findings support **Hypothesis 4**, we conduct a correlation analysis between the participants' stylistic tendencies and the average performance difference between MIND MELD and DAgger. We find that participants' delayed/anticipatory tendencies significantly correlate with MIND MELD's advantage over DAgger $(r(40) = .36, p = .017)$, as shown in Figure 3.6.

We lastly investigate our findings in the context of **Hypothesis 5** to determine if MIND MELD is rated subjectively higher by participants. We conducted a repeated measures ANOVA with a Tukey post hoc or Friedman's test (see omnibus statistics in the Appendix). As shown in Table B.2, MIND MELD is rated statistically significantly higher compared to both DAgger and BC for all subjective metrics. These findings support **Hypothesis 5**.

3.5.1 Sensitivity Analysis for Ground Truth Labels

To determine how close to optimal the ground truths need to be for MIND MELD to outperform DAgger, we conduct a sensitivity analysis. We train MIND MELD on ground truths which are incrementally shifted from optimal and compare the performance of the agent that has been trained via MIND MELD using suboptimal ground truths to DAgger's performance.

We expect to see that, as the ground truth labels are shifted further from the optimal, MIND MELD's advantage over DAgger in terms of ability to reach the goal will decrease. Fig. Figure 3.7 shows a plot of the percentage by which the ground truths are shifted from the optimal versus MIND MELD's advantage over DAgger in terms of average distance from the goal. We find that MIND MELD's advantage over DAgger negatively correlates with the amount by which the ground truths are shifted $(r(3) = -0.97, p = .0072)$ and that

MIND MELD outperforms DAgger even if the ground truths have been shifted by as much as 15% from the optimal. These results suggest that MIND MELD is robust to deviation of the ground truth labels from optimal.

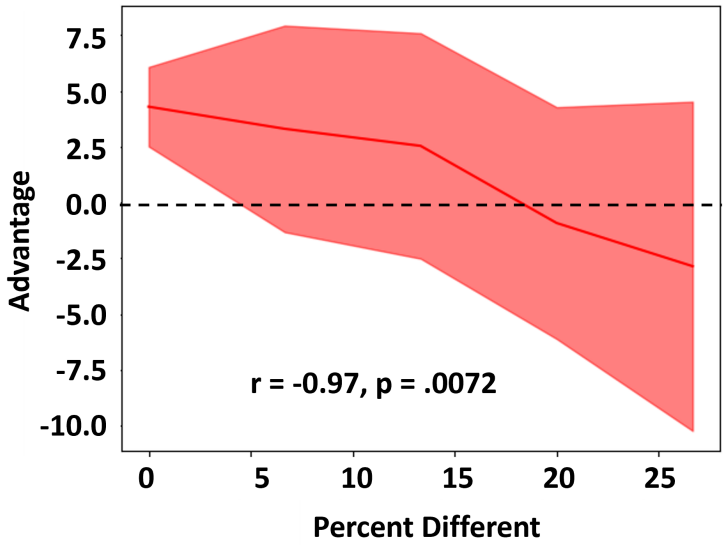


Figure 3.7: This figure shows the percentage by which the ground truths deviate from optimal versus the advantage that MIND MELD has over DAgger.

3.5.2 Importance of Personalized Embeddings

In this section, we conduct an analysis to determine if personalization of the embeddings improves the ability of the agent to learn from the demonstrator. To investigate the importance of the personalized embeddings, for each subject, we randomly select a personalized embedding belonging to another subject and use this random embedding to map the subject's corrective labels to new labels. Then, we retrain the agent on the adjusted, corrective labels. We hypothesize that agents trained on corrective labels that have been mapped using a random personalized embedding will perform worse than agents trained on corrective labels mapped using the embedding that was learned for the specific participant. To investigate our hypothesis, we conduct a Friedman's test to determine if there is a statistically significant difference between the average distance to the goal for agents trained on the correct embedding versus the random embedding. We find that the average distance from the goal is statistically

significantly larger ($\chi^2(1) = 9.92, p = .0016$) when a random embedding is used compared ($M = 22.2, SD = 7.1$) to when the correct embedding is used ($M = 20.4, SD = 5.5$). We also employ Spearman's test to determine if the amount by which the random embedding differs from the original embedding correlates with decreased performance. We find a significant correlation between the percent increase in distance from the goal and the distance between the random embedding and original embedding ($\rho = 0.28$, $p < .001$). These results suggest that the personalization of embeddings is an important factor contributing to the success of MIND MELD.

3.6 Discussion

In our analysis, we find support for **Hypotheses 1-5**, illustrating that MIND MELD can learn stylistic tendencies of suboptimal and heterogeneous demonstrators, map the suboptimal demonstrations to better demonstrations, and, as a result, outperform prior work in both robot-centric and human-centric LfD. We find that MIND MELD is able to learn various participant styles, such as participants' tendency to over-/under-correct ($p < .001$) and provide delayed and anticipatory demonstrations ($p < .001$), suggesting that MIND MELD can provide positive results with a diverse user pool. For more discussion on stylistic tendencies, please refer to the Appendix.

Because MIND MELD is able to learn heterogeneous tendencies and utilize this information to correct for suboptimal behavior, we find that MIND MELD outperforms prior work in terms of its ability to reach the goal in an LfD task. MIND MELD achieves both a higher probability of reaching the goal and a lower average distance from the goal compared to both baselines, DAgger ($p < .001$) and BC ($p < .001$). Additionally, we observe that the more delayed a participant is at providing demonstrations, the better MIND MELD performs over DAgger ($p = .017$). We find that, for participants who provide less suboptimal demonstrations, MIND MELD and DAgger exhibit more similar performance because there is less of a need to correct a participants' demonstrations. When a participant's behavior

deviates more from the optimal, DAgger performs worse, whereas MIND MELD is able to correct for the suboptimality.

Not only do we see improved performance in terms of objective metrics, we also find that MIND MELD outperforms both DAgger and BC in terms of our subjective metrics. Participants rate MIND MELD to be more likeable ($p = .004$), intelligent ($p = .008$), and trustworthy ($p = .001$) compared to DAgger. Additionally, we find the participants' perceived workload is rated as lower for MIND MELD ($p = .005$). This is an interesting finding considering that for both MIND MELD and DAgger, participants are tasked with providing corrective demonstrations to the agent. With respect to performance and human usability, MIND MELD achieves the best of both worlds. MIND MELD improves upon the performance of robot-centric algorithms, while being easy to teach, likeable, intelligent, and trustworthy.

3.7 Limitations/Future Work

Due, in part, to the recruiting difficulties imposed by the COVID-19 pandemic, our sample population consisted primarily of students with a mean age of 22.6. In the future, we plan to conduct this experiment with a more diverse set of participants. We also note that MIND MELD requires training participants to meta-learn the model parameters and a set of calibration tasks with ground-truth labels to learn the personalized embeddings. However, our results demonstrate that MIND MELD improves the quality of the corrective demonstrations by 37.6% and LfD outcomes ($p < .001$), making this additional step worthwhile.

Additionally, MIND MELD makes several assumptions, listed in Section 3.2, about the way in which individuals provide corrective demonstrations. Yet, the success of our algorithm suggests that these assumptions appear to be sufficiently met for our experimental setup. For this study, we assume that a person's demonstration style will remain constant; however, we do expect that, over a longer period of interaction, a person's style of demon-

strations may change and adapt. In future work, we plan to investigate how to update our framework to account for and learn changing styles during longitudinal LfD.

Lastly, we aim to investigate if we can replicate the benefits of MIND MELD in other domains. We plan to implement MIND MELD on a robot arm domain, which may produce different behavior and stylistic tendencies amongst participants due to more degrees of freedom and a more complex user interface.

3.8 Conclusion

In this chapter, I introduce MIND MELD, a novel LfD framework that learns personalized embeddings for heterogeneous demonstrators and improves upon suboptimal human feedback for robot-centric LfD algorithms. Through a human-subjects experiment, we showed that MIND MELD outperforms a human-centric baseline, BC, and a robot-centric baseline, DAgger, with regards to multiple measures of algorithm performance. Furthermore, users found MIND MELD more intelligent, likeable, trustworthy, and easier to teach than BC and DAgger.

Our MIND MELD framework fills a gap in prior work by introducing an approach which enables a robot to learn from suboptimal and heterogeneous demonstrators in a robot-centric LfD paradigm. This approach exemplifies the potential for personalized frameworks to improve human-machine interaction. In keeping with the goals of my thesis, I have demonstrated that MIND MELD enables learning from a large population while also personalizing for an individual end-user. By accounting for end-user heterogeneity, we are able to improve upon the human-machine relationship as demonstrated in a large human-subjects study.

CHAPTER 4
PERSONALIZED TEACHING VIA RECIPROCAL MUTUAL INFORMATION DRIVEN META-LEARNING FROM DEMONSTRATION

4.1 Introduction

In the previous Chapter, I introduced MIND MELD, a personalized framework for learning from suboptimal and heterogenous demonstrators. I demonstrated that MIND MELD is capable of effectively learning a personalized embedding describing a demonstrator's style and mapping the demonstrator's suboptimal demonstrations to better demonstrations. Despite MIND MELD's positive results, by correcting for suboptimality under-the-hood, the robot and human will not have a shared understanding about how best to accomplish the task. This lack of transparency can lead to decreased trust in the system and a lower likelihood of task success [101, 119]. Instead, in this chapter, I propose an approach to provide personalized robotic feedback to demonstrator's to improve upon their suboptimality, thus increasing transparency and creating a shared understanding between the human and robot.

Many non-expert users lack a functional understanding of the robotic systems they are teaching, which may contribute to their suboptimal tendencies. This lack of understanding is concerning because prior work has shown that trust and reliance decrease when the end-user does not understand how the robot operates [101, 119]. Furthermore, if an underlying algorithm is correcting for teacher suboptimality, the teacher will likely never learn to be a better demonstrator. Correcting for suboptimality via an algorithm without communicating to the demonstrator how to improve upon their demonstrations is problematic for several reasons. Producing positive results from poor demonstrations will only reinforce the teacher's suboptimal tendencies, thereby preventing the teacher from improving. Additionally, re-

inforcing low quality and suboptimal demonstrator tendencies will likely have long-term consequences. For example, when the teacher provides demonstrations to a different robotic platform that may be incapable of correcting for demonstrator suboptimality, the robot will struggle to learn from the teacher.

Furthermore, as shown in prior work, humans may generalize better to out-of-distribution tasks and, if they are capable of providing high-quality demonstrations, will be more effective teachers [28]. Therefore, we hypothesize that correcting for suboptimal demonstrations under-the-hood as MIND MELD does may not be the best long-term strategy because doing so may 1) contribute to end-users' lack of functional understanding, 2) reinforce suboptimal tendencies, and 3) result in poor performance on out-of-distribution tasks and novel robotic platforms [101, 119, 28]. Consequently, there is a need for a framework that can coach demonstrators to become better teachers.

While several approaches have attempted to improve upon a teacher's ability to provide high quality demonstrations via tutorials and videos [60, 59, 120], prior work has primarily focused on correcting for suboptimality after-the-fact rather than directly improving teaching abilities. In Chen et al., the authors introduce Self-Supervised Reward Regression (SSRR) in which the authors improve upon an agent's ability to learn from suboptimal demonstrations by characterizing the relationship between noise and performance [41]. Their approach bootstraps off of suboptimal demonstrations to learn an idealized reward function. Similarly, T-Rex and D-Rex improve upon the ability to learn from suboptimal demonstrations by learning a reward function from a ranked set of demonstrations [37, 16].

Rather than correcting for suboptimality under-the-hood as much of prior work does, in this chapter, I introduce an approach to provide personalized robotic feedback to a human demonstrator to directly improve upon their ability to provide high-quaility demonstrations, To achieve this goal, I propose Reciprocal MIND MELD (Reciprocal Mutual Information Driven Meta-Learning from Demonstration (MIND MELD)). Reciprocal MIND MELD is based upon the MIND MELD framework but is meant to guide the human demonstrator to

proactively improve their feedback. Reciprocal MIND MELD differs from MIND MELD in three significant ways. First, Reciprocal MIND MELD learns a *semantically meaningful* personalized embedding via calibration tasks that describes the way in which a demonstrator is suboptimal (Figure 4.1b and e). Second, based upon this personalized embedding, Reciprocal MIND MELD provides *robotic feedback* to the demonstrator to improve their teaching abilities (Figure 4.1c and f) and consequently improve learning outcomes for the agent rather than correcting for suboptimality retroactively. Third, we introduce an Embedding Predictor Network (EPN) which dynamically updates the demonstrator's personalized embedding by estimating its new location (Fig. 1e), thus eliminating the need to repeat the calibration tasks. In this chapter, we contribute the following:

1. We propose Reciprocal MIND MELD, a novel method for providing personalized feedback to demonstrators to improve their teaching abilities via a personalized embedding.

2. We develop an EPN to dynamically update the demonstrator's personalized embedding and without the need to repeat the time-consuming calibration tasks (Figure 4.1).

3. We demonstrate that Reciprocal MIND MELD can improve an individual's demonstrations ($p < .001$), accurately estimate a demonstrator's new embedding ($p = .002$), and improve learning outcomes of the robot ($p = .045$) in a driving simulator domain.

4.2 Preliminaries

Reciprocal MIND MELD is inspired by the MIND MELD architecture demonstrated in previous work [1]. The objective of MIND MELD is to learn a personalized embedding to describe the way in which a demonstrator is suboptimal in an Robot-Centric (RC) LfD paradigm in which the demonstrator provides corrective feedback to the robot. MIND MELD then utilizes this embedding to map a demonstrator's suboptimal demonstrations

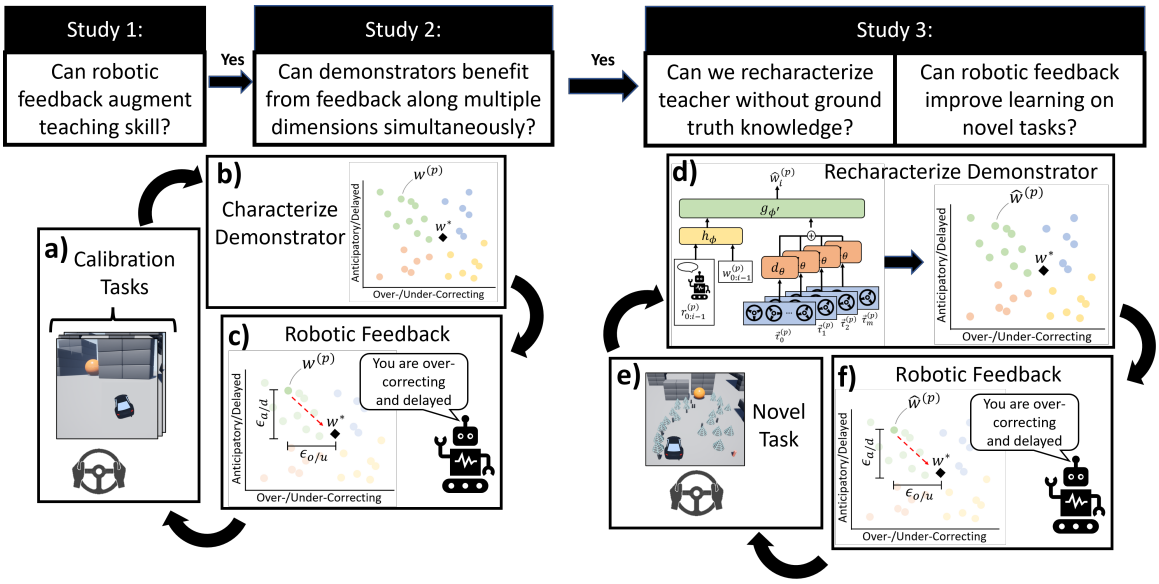


Figure 4.1: This figure illustrates an overview of our methodology and study designs. Figs 1a, 1b, and 1c show the methodology for Studies 1 and 2 and Figs 1d, 1e, and 1f the methodology for Study 3.

to demonstrations closer to the optimal. The MIND MELD architecture (shown in gray in Figure 4.2) is trained via *calibration tasks*, which are used to learn the mapping (f_θ, $\mathcal{E}_{\phi'}$, and q_ϕ) from suboptimal labels, $a^{(p)}_{t-\Delta t:t+\Delta t}$, to better labels, $\hat{d}^{(p)}_t$, and learn the personalized embedding, $w^{(p)}$, representing an individual demonstrator. The calibration tasks consist of a set of pre-recorded policy rollouts with known optimal labels. Participants provide corrective demonstrations to the robot during these rollouts to direct the robot to a goal. MIND MELD learns to map the participant's corrective labels to higher-quality labels while simultaneously inferring the personalized embedding, $w^{(p)}$, representing an individual, p's, suboptimal style. To ensure that $w^{(p)}$ can represent various and distinct feedback styles, MIND MELD maximizes a lower bound on mutual information between the way in which a demonstrator is suboptimal and $w^{(p)}$ via variational inference [108]. Additional details can be found in the Appendix.

While prior work demonstrated that MIND MELD is capable of improving upon suboptimal demonstrations, MIND MELD suffers from several key limitations: 1) MIND MELD corrects for suboptimality under-the-hood and does not convey to the demonstrator how best

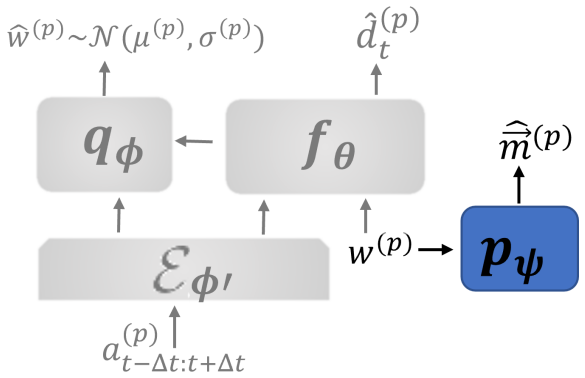


Figure 4.2: This figure shows the MIND MELD architecture from Schrum et al. [2] in gray and the additional network head, p_ψ, in blue for learning a semantically meaningful embedding space (see Subsection 4.3.1).

to improve their suboptimal tendencies, and 2) MIND MELD assumes that demonstrators are static (i.e., the way in which they are suboptimal does not change over time). Reciprocal MIND MELD overcomes these limitations by 1) providing actionable robotic feedback to the demonstrator to improve upon the quality of their demonstrations and 2) dynamically updating the estimate of their personalized embedding online via our EPN in order to account for changes in suboptimal tendencies and teaching ability.

Driving Simulator Domain - In keeping with prior work [2], we utilize a driving simulator domain based on the high-fidelity physics simulator, Airsim, and an Xbox steering wheel to evaluate Reciprocal MIND MELD. Driving simulators allow researchers to study novel algorithms in an environment that is safe for human subjects. In this domain, participants are tasked with teaching a car to drive from a start location to a goal in various environments while avoiding obstacles. The action space consists of the position of the wheel (-540° to 540°), and the state space consists of images, position, velocity, and acceleration. Feedback is provided to demonstrators via verbal instructions.

4.3 Methodology

Because humans have a greater ability to generalize to novel tasks and domains than a machine-learning algorithm [28], our objective is to provide demonstrators with knowledge

about how to improve their demonstrations rather than correcting suboptimality under-the-hood. We propose an approach to reason about a demonstrator's embedding and provide robotic feedback derived from their embedding that is intended to improve upon their demonstration abilities. In keeping with prior work [2], we investigate the abilities of our approach in a driving simulator domain. We break the problem of improving upon a demonstrator's teaching abilities into three research questions.

RQ1: Can robotic feedback improve upon a demonstrator's teaching abilities?

RQ2: What is the best method to provide robotic feedback to improve teaching abilities?

RQ3: Does robotic feedback result in improved learning outcomes on novel tasks and over time?

4.3.1 Semantically Meaningful Embedding Space

Prior work [1, 2] has illustrated that MIND MELD learns embeddings that correlate with suboptimal tendencies and that demonstrators tend to over-/under-correct and provide anticipatory/delayed feedback in a driving simulator domain. We note that domain expertise is required to determine these dimensions of suboptimality. This suboptimality is related to the unintuitive nature of RC LfD as well as the correspondence problem [38, 121] which arises from differences in embodiment between humans and robots. These suboptimal tendencies are unrelated to the specific task itself, but are related to the task specifications (e.g., providing corrective feedback via a steering wheel). While there may be additional dimensions of suboptimality depending on the robotic domain, we focus our investigation on the over-/under- (o/u) and anticipatory-/delayed- (a/d) dimensions, as these were determined in prior work to be principle dimensions of suboptimality [2]. We posit that these two dimensions will be common across RC LfD paradigms which require continuous control input. We aim to test this hypothesis in future work. Our goal is to learn a *semantically meaningful* embedding space (i.e., a space that can be translated into robotic feedback) and then utilize the location of the demonstrator's embedding within the embedding space to

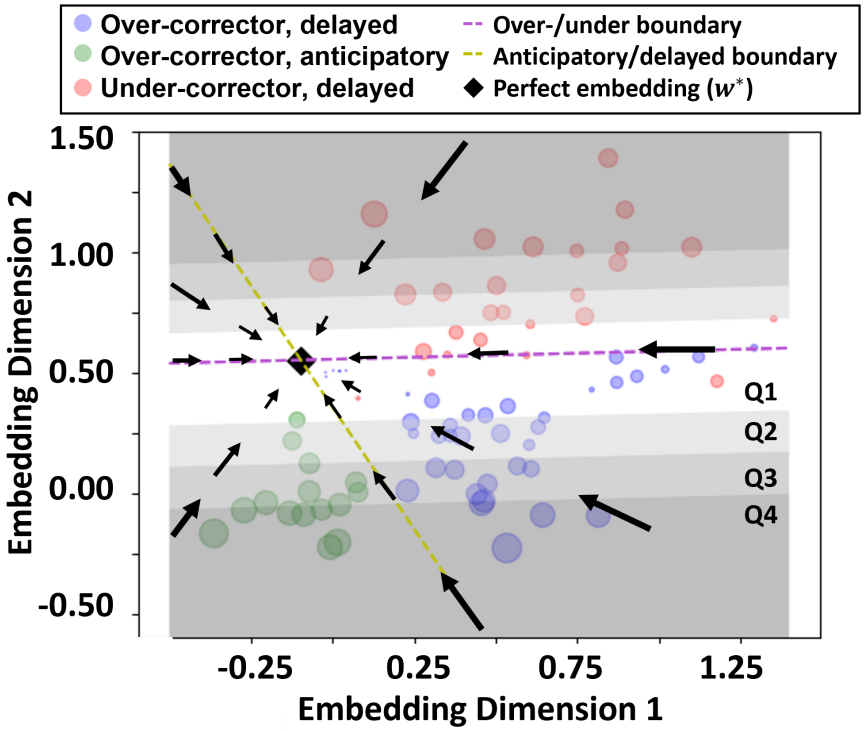


Figure 4.3: The learned embedding space and decision boundaries. Each point represents the embedding of a demonstrator, and the diameter represents the magnitude of over-/under-correction. The arrows indicate the direction an embedding should move to be closer to the perfect embedding. A similar plot showing the magnitude and quartiles for the anticipatory/delayed dimension can be found in the Appendix. Blue points represent participants who tend to over-correct and are delayed, red points represent participants who under-correct and are delayed, and green represent those who over-correct and are anticipatory. The yellow line represents the decision boundary for the a/d dimension and the purple line represents the boundary for the o/u dimension. This plot demonstrates that demonstrators that are similar in the way in which they are suboptimal tend to cluster together and these these suboptimal tendencies are linearly separable.

provide actionable robotic feedback.

To learn a semantically meaningful embedding space whose dimensions reflect suboptimal tendencies, we add an additional network head, $p_\psi(w^{(p)}) = \hat{\vec{m}}^{(p)}$, as shown in blue in Figure 4.2, to the MIND MELD architecture to estimate the suboptimal tendency, $\vec{m}^{(p)}$, (i.e., the magnitude by which the demonstrator over-/under-corrects and is anticipatory/delayed). We utilize a Mean-squared error (MSE) loss, $L(\psi, w) = \frac{1}{N}\sum_i ||p_\psi(w^{(i)}) - \vec{m}^{(i)}||_2^2$, to train the network to predict the suboptimal tendency, $\vec{m}^{(p)}$, given the personalized embedding.

This loss helps to ensure that the dimensions of the embedding space are semantically meaningful and can therefore be translated into actionable robotic feedback. Under IRB approval, we leverage the calibration dataset collected in Schrum et al. [2] to learn a semantically meaningful embedding space. This dataset consists of 76 participants who provided demonstrations on a set of calibration tasks. The suboptimal magnitude, $\vec{m}^{(p)}$, is determined via Dynamic Time Warping (DTW) [114] between the participants' feedback and optimal labels from the calibration tasks. Because MIND MELD outputs the difference between the participant's corrective label and the optimal label, the perfect demonstrator's embedding, w^*, is defined as the embedding which minimizes the output of the MIND MELD architecture as shown in Equation 4.1. Here, $a^{(p)}_{t-\Delta t:t+\Delta t}$ is a sequence of demonstrations.

$$w^* = \underset{w^{(p)}}{\operatorname{argmin}} \sum_{t,p} \mathcal{M}_{\theta,\phi'}(a^{(p)}_{t-\Delta t:t+\Delta t}, w^{(p)}) \tag{4.1}$$

Our next objective is to determine the semantically meaningful dimensions of the embedding space. We train a Support Vector Machine (SVM) with a linear kernel to learn the decision boundaries which best separate the demonstrators into their respective suboptimal categories (o/u and a/d). The SVM training labels are determined via DTW between the participant labels and the optimal labels from the calibration tasks. We utilize an SVM to learn the decision boundaries so that we can add the additional constraint that the classifier must pass through the point representing the perfect demonstrator, w^*. The distance between the embedding and the decision boundary along the suboptimal dimension determines the magnitude by which the demonstrator is suboptimal.

Figure 4.3 depicts our embedding space with linear classifiers separating over-correctors from under-correctors and delayed from anticipatory. The size of the point represents the magnitude by which the demonstrator is suboptimal in the o/u dimension as determined by DTW. The plot illustrates that demonstrators who are more suboptimal in o/u (as represented by larger points) are farther from the o/u decision boundary, supporting our hypothesis that distance from the decision boundary can be used to measure the degree of suboptimality.

To further support our claim, we apply Spearman's correlation and find that distance from the decision boundary strongly correlates with magnitude of suboptimality in both the o/u ($\rho = .84$, $p < .001$) and in a/d dimensions ($\rho = .93$, $p < .001$).

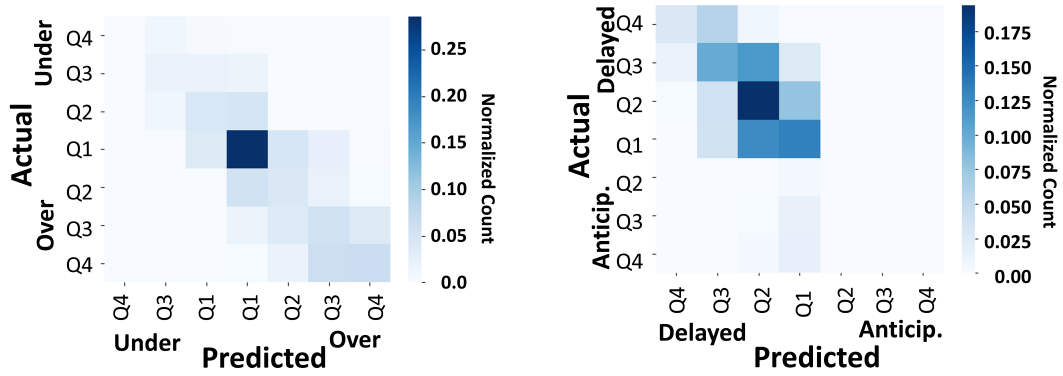
4.3.2 Robotic Feedback

To determine the feedback the robot should provide, we calculate the distance, ϵ, along the semantically meaningful dimension between the personalized embedding, $w^{(p)}$, and perfect embedding, w^*, as shown in Figure 4.1. In the driving domain, we are interested in $\epsilon^{(i)}_{o/u}$ and $\epsilon^{(i)}_{a/d}$, which define the distance between the demonstrator's embedding and the hypothetical perfect demonstrator's embedding in the o/u dimension and the a/d dimension respectively after the i^{th} round of feedback. In our framework, the feedback is proportional to the distance from w^*.

Table 4.1: This table shows the feedback a participant receives based on their quartile and study condition for Study 1. Analogous feedback for the Cooperative condition is provided in Study 2 for the anticipatory/delayed dimension in addition to the over-/under-correcting dimension.

Cooperative Quartile	**Adversarial Quartile**	**Robotic Feedback**
First	Fourth	"Your feedback is good! Keep it up."
Second	Third	"You are slightly over-/under-correcting. Please turn the wheel a bit more/less."
Third	Second	"You are over-/under-correcting. Please turn the wheel more/less."
Fourth	First	"You are over-/under-correcting a lot. Please turn the wheel a lot more/less."

To convert $\epsilon^{(i)}_{o/u}$ into actionable and intelligible robotic feedback, we discretize the range of $\epsilon^{(i)}_{o/u}$ by splitting the embeddings from the previously collected calibration participants into quartiles as shown in Figure 4.3. Our objective is to move a participant's embedding so that they are in the range denoting the 25% of calibration participants who are the least suboptimal (i.e., quartile one). Participants who fall in a quartile farther from the decision



(a) Over-/under-correcting

(b) Anticipatory/delayed

Figure 4.4: Figure 4.4(a) and Figure 4.4(b) show the confusion matrices for for predicting the quartile that the embedding falls within on holdout test tasks.

boundary receive feedback proportional to their quartile. For example, if a participant falls in the fourth quartile in the o/u dimension, the robot will instruct the participant to turn the wheel a lot less compared to slightly less in the second quartile. A table showing the feedback for each quartile and dimension can be found in Table Table B.1.

4.3.3 Online Embedding Estimate

To determine if additional feedback should be provided to the demonstrator and if so, the form of the feedback, we must update our estimate of $w^{(p)}$ after each iteration of robotic feedback. One option to update our estimate of the embedding is to have the demonstrator redo the calibration tasks. However, doing so is time consuming and increases the workload of the demonstrator.

Instead, we propose to dynamically update the embedding online using a Long-Short Term Memory Network (LSTM)-based architecture which extracts salient features from the demonstrations to estimate the personalized embedding rather than relying on calibration tasks which require known, optimal labels. For example, the velocity and magnitude with which the demonstrator turns the steering wheel are two salient features which can inform

0.50

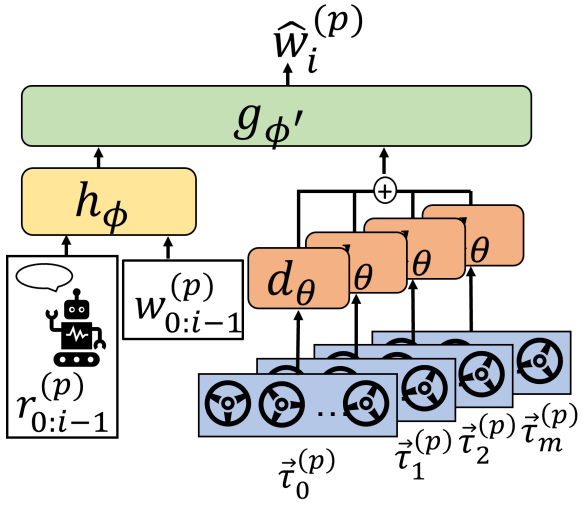


Figure 4.5: This figure illustrates our EPN architecture. $\tau_{0:m}^{(p)}$ is the set of demonstrations provided by the participant in round i. $r_{0:i-1}^{(p)}$ is the set of previous robotic feedback provided to the demonstrator and $w_{0:i-1}^{(p)}$ are the participant's previous embeddings.

the estimate of the new embedding. We call this network the EPN (Figure 4.5). The input to the EPN is the set of new demonstrations, $\tau_{0:m}^{(p)}$, the demonstrator's previous embeddings, $w_{0:i-1}^{(p)}$, and the robotic feedback that was previously provided to the demonstrator, $r_{0:i-1}^{(p)}$. The output of the EPN is an estimate of the new personalized embedding, $\hat{w}_i^{(p)}$. This network utilizes two LSTM subnetworks, h_ϕ and d_θ, the output of which is then fed into subnetwork, $g_{\phi'}$, made up of linear layers with ReLU activations. The inputs to h_ϕ are $w_{0:i-1}^{(p)}$ and $r_{0:i-1}^{(p)}$. Each trajectory, $\tau_t^{(p)}$, is fed into an LSTM subnetwork, d_θ. We then average across the outputs of d_θ and feed the result into $g_{\phi'}$ which produces our embedding estimate, $w_i^{(p)}$. We choose to average across the outputs of d_θ so that our network is agnostic to the number of trajectory inputs.

We train our EPN on the data collected in Studies 1 and 2 as described in Section 4.4. Figure 4.4(a) and Figure 4.4(b) show confusion matrices depicting the ability of the network to accurately predict the quartile of suboptimality in the o/u dimension and the a/d dimension respectively on holdout test tasks.

4.4 Human-Subjects Studies, Results, and Discussion

To determine if Reciprocal MIND MELD is able to improve upon a demonstrator's ability to provide high-quality demonstrations, we conduct three human-subjects studies. The objective of Study 1 is to determine if we are able to shift a demonstrator's embedding via verbal robotic feedback in the o/u dimension (R1). In Study 2, we investigate if, and how best, we can shift a demonstrator's embedding in two dimensions (R2). In Study 3, we determine if 1) robotic feedback derived from our EPN rather than the calibration tasks is a good metric of teacher suboptimality and 2) if robotic feedback improves teaching outcomes over time (RQ3). During each study, we measured trust, team fluency, workload, and understanding to determine how robotic feedback altered participants' subjective attitude towards each agent. In our analysis, we check parametric models for normality and homoscedasticity. Model details, tests for assumptions, and additional results are in the Appendix. Each study is approved by Georgia Tech's IRB

Table 4.2: This table shows the mean, (standard deviation), and test statistics of the subjective metrics and $\Delta\epsilon_{o/u}$ for Study 1. Δ Trust and Δ Fluency describe the change in Trust and Fluency respectively between rounds one and four.

	Cooperative	**Adversarial**	**None**	Test Statistic	p-value
$\Delta\epsilon_{o/u}$	0.33 (0.2)	-0.30 (0.2)	0.01 (0.2)	$F(2, 24) = 20.2$	$p < .001$
Workload	37.5 (16.4)	46.1 (19.5)	53.5 (11.6)	$F(2, 24) = 2.21$	$p = .132$
Likeability	6.69 (2.0)	6.81 (1.5)	6.86 (1.4)	$F(2, 24) = .024$	$p = .978$
Intelligence	6.31 (1.6)	5.57 (1.1)	6.24 (1.4)	$F(2, 24) = 1.03$	$p = .372$
Δ Trust	0.56 (0.4)	-0.01 (0.4)	0.05 (0.2)	$F(2, 24) = 5.15$	$p = .014$
Δ Fluency	0.34 (0.4)	-0.13 (0.3)	-0.04 (0.4)	$F(2, 24) = 5.10$	$p = .014$

4.4.1 Study 1 (RQ1)

Our objective in Study 1 is to demonstrate that robotic feedback can effectively modulate a participant's teaching. In this study, we start by investigating feedback only in the o/u dimension. After completing pre-study surveys, participants complete four rounds of the calibration tasks to measure how their embedding is changing. Participants receive

robotic feedback between each round and complete trust [115] and fluency [122] surveys to determine their subjective perceptions of the robot.

0.50

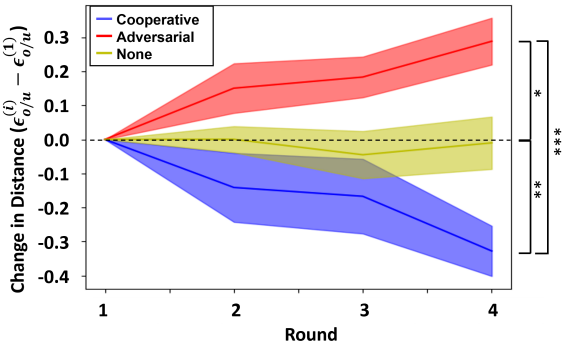


Figure 4.6: This figure shows the difference between the embedding distance at round i, $\epsilon^{(i)}_{o/u}$, and the embedding distance at round one, $\epsilon^{(1)}_{o/u}$, in the o/u dimension for Study 1.

Conditions: In the *Cooperative* condition, the robot provides feedback to improve the demonstrator's teaching. In the *Adversarial* condition, the robot provides feedback to make the participant a worse demonstrator. We include this condition to determine if we are able to move a participant's direction in *any* direction and to ensure that a participant's abilities are not changing simply due to interaction with the agent. In the *None* condition, the participant does not receive any feedback.

Results: In Study 1, we recruited 27 participants (Mean age = 24.15, SD = 3.4; 37.0% Female). Figure 4.6 shows the change in the distance ($\epsilon^{(i)}_{o/u} - \epsilon^{(1)}_{o/u}$) in the o/u dimension between round one and rounds one through four. We plot $\epsilon^{(i)}_{o/u} - \epsilon^{(1)}_{o/u}$ to show how participants change irrespective of their initial teaching skill. We find that the distance at round one, $\epsilon^{(1)}_{o/u}$, is significantly greater from the distance, $\epsilon^{(4)}_{o/u}$, at round four in Cooperative ($\chi^2(1) = 5.44$, $p = .020$) suggesting that participants' embeddings move closer to perfect embedding. Adversarial produces the opposite result, i.e. the embedding moves significantly farther from the perfect embedding ($F(1, 8) = 20.1$, $p = .002$).

We additionally find that Adversarial results in the embedding shifting significantly farther from the perfect embedding between rounds one to four ($F(2, 24) = 20.2$, $p < .001$)

compared to Cooperative ($p < .001$) and None ($p = .014$). Cooperative shifts the embedding significantly closer to the perfect embedding ($p = .009$) compared to None. Together, these findings indicate that our approach is capable of modulating teaching style in either direction along the suboptimal dimension. Further, the results in None shows that participants are not simply improving due to repeated interactions.

0.50

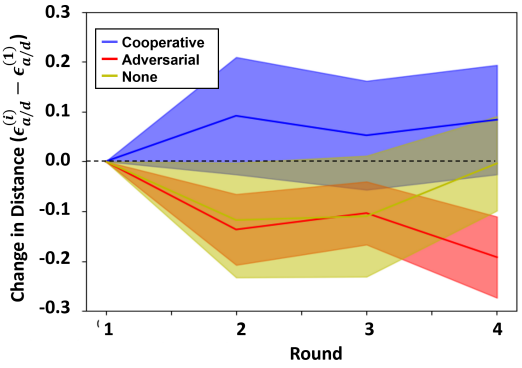


Figure 4.7: This figure shows the difference between the embedding distance at each round, $\epsilon_{a/d}$, and the embedding distance at round one, $\epsilon_{a/d}^{(1)}$ for the a/d dimension.

Interestingly, we find that participants become significantly worse in the a/d dimension when they only receive Cooperative feedback in the o/u dimension. Participants become better in the a/d dimension when they receive Adversarial feedback in the o/u dimension. Figure B.4(c) shows the change in the amount by which a participant provides anticipatory/delayed feedback as calculated by the distance from the perfect demonstrator in embedding space. We show that as participants improve in the over-/under-correcting dimension, they tend to become worse in the anticipatory/delayed dimension and vice versa when no feedback is provided. This suggests that the task of improving participants demonstration quality in both the over-/under-correcting dimension and the anticipatory/delayed dimension may be particularly difficult since improving in the over-/under-correcting dimension tends to produce greater suboptimality in the anticipatory/delayed dimension.

We find that participants' trust (Table B.2) increased significantly more ($F(2, 24) = 5.15$, $p = .014$) in Cooperative compared to Adversarial ($p = .020$) and None ($p =$

.038). Additionally, we find a positive change in fluency ($F(2, 24) = 5.10$, $p = .014$) in Cooperative compared to Adversarial ($p = .017$) as shown in Table B.2. **Takeaway: Robotic feedback can effectively improve a participant's teaching abilities in a driving simulator domain**.

4.4.2 Study 2 (RQ2)

0.50

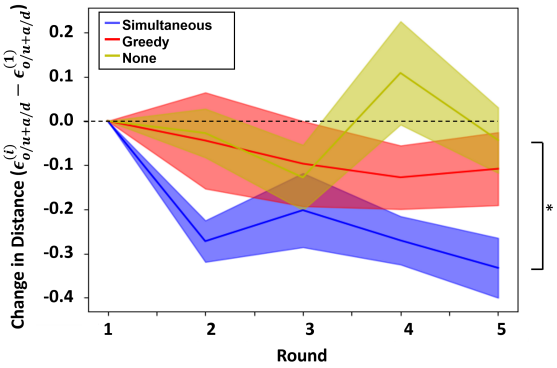


Figure 4.8: This figure shows the difference between the embedding distance at round i, and the embedding distance at round one for Study 2.

In Study 2, we next determine how best to provide robotic feedback to both prevent cognitive overload and efficiently improve upon a participant's teaching abilities. Our study design follows the same procedure as Study 1, in which participants complete five rounds of the calibration tasks, receive robotic feedback between each round, and complete surveys between each round.

Conditions: In *Simultaneous*, the robot provides feedback related to both the o/u and the a/d dimensions. In *Greedy*, the robot only provides feedback related to the condition in which the participant is worst (i.e., furthest from w^*). This condition is intended to reduce cognitive overload for the demonstrator as it reduces the amount of information provided at once. In *None*, the participant receives no feedback.

Results: We recruited 39 participants (Mean age = 22.46, SD = 3.3; 38.5% Female). Figure B.5 shows the overall change in the distance ($\epsilon^{(i)}_{o/u+a/d} - \epsilon^{(1)}_{o/u+a/d}$) in the two dimensions

of suboptimality between round one and rounds one through five. We find that the distance at round one, $\epsilon^{(1)}_{o/u+a/d}$, is significantly greater from the distance, $\epsilon^{(5)}_{o/u+a/d}$, at round five in Simultaneous ($F(1, 12) = 22.3$), $p < .001$). We next compare $\Delta\epsilon_{o/u+a/d}$ across conditions ($F(2, 36) = 3.77$, $p = .033$). We find that Simultaneous results in the embedding shifting significantly closer to the perfect embedding between rounds one to five compared to None ($p = .034$). We do not find significance between None and Greedy or Simultaneous and Greedy. We additionally find that Simultaneous results in an improvement across both the o/u and a/d dimensions of suboptimality.

Additionally, we find that participants' trust ($F(2, 36) = 3.81$, $p = .032$) and team fluency ($F(2, 36) = 7.23$, $p = .002$) significantly increased in Simultaneous compared to None ($p = .029$, $p = .002$ respectively). Lastly, although the result is not significant, we find that participant's understanding of the robot increased more in the Simultaneous ($M = 0.61$, $SD = 0.62$) condition compared to None ($M = 0.14$, $SD = 0.69$) and Greedy ($M = .15$, $SD = 0.58$). **Takeaway: Providing feedback in both dimensions simultaneously produces better results for both objective and subjective metrics.**

4.4.3 Study 3 (RQ3)

0.50

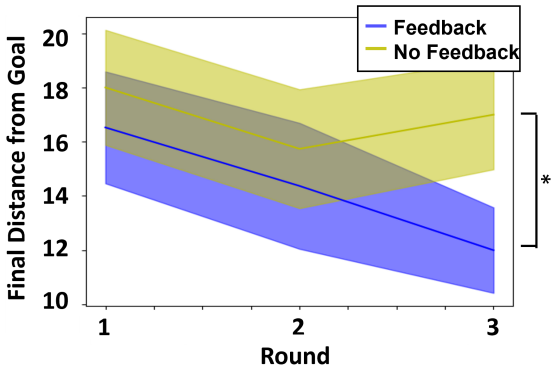


Figure 4.9: This figure shows the final distance from the goal for the robot after each round of Study 3.

In Study 3, we aim to show that our approach and the results from Study 1 and 2 translate

to improved learning outcomes for an LfD agent on novel tasks. Participants first complete the calibration tasks to obtain an initial estimate of their embedding, $w_0^{(p)}$, and determine $\epsilon_{o/u}^{(0)}$ and $\epsilon_{a/d}^{(0)}$. Next, the robot provides feedback to the participant intended to improve their demonstrations in both the o/u and the a/d dimension given our positive findings for the *Simultaneous* condition in Study 2. Participants then train the robot for three rounds in three different novel environments (i.e., new start and goal locations) for six demonstrations each. Between each environment, we estimate the participant's new embedding, $w_i^{(p)}$, via the EPN, and calculate $\epsilon_{o/u}^{(i)}$ and $\epsilon_{a/d}^{(i)}$ after each round, $i \in \{1, 2, 3\}$. The robot provides robotic feedback based upon the new estimate of the participant's embedding derived from the EPN. At the end of the study, the participants redo the calibration tasks to determine $\epsilon_{o/u}^{(4)}$ and $\epsilon_{a/d}^{(4)}$. By redoing the calibration tasks, we are able to obtain a ground truth estimate of how the quality of their demonstrations has changed over the course of the study.

Conditions: In *Feedback*, the robot provides feedback to the participant about their demonstrations. In *No Feedback*, the robot still interacts with the participant but does not provide feedback.

Results: We recruited 60 participants (Mean age = 21.9, SD = 2.89; 28.3% Female). Figure 4.9 shows the robot's final distance from the goal for Feedback and No Feedback for rounds 1-3. Participants in Feedback achieve a lower final distance to the goal in round one despite starting off as worse demonstrators on average as measured via the initial calibration tasks (Mean $\epsilon_{o/u+a/d}^{(0)}$ in Feedback: 0.93, Mean $\epsilon_{o/u+a/d}^{(0)}$ in No Feedback: 0.89). Additionally, the final distance of the robot improves over the rounds in Feedback whereas in No Feedback, the robot improves slightly then gets worse in the final round. In round three, the robot achieves a significantly lower final distance from the goal ($Z = -2.0$, $p = .045$) compared to No Feedback.

To determine if the embedding as estimated by the EPN is a good metric of performance, we compute the correlation between the distance, $\epsilon_{o/u+a/d}^{(i)}$, of the estimated embedding from the perfect embedding and performance as measure by the average distance from the goal

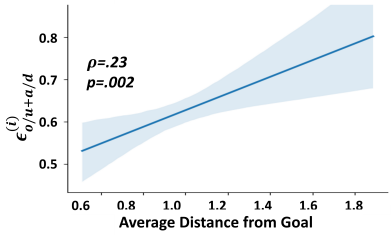


Figure 4.10: Correlation between the embedding distance, $\epsilon^{(i)}_{o/u+a/d}$, and distance from the goal.

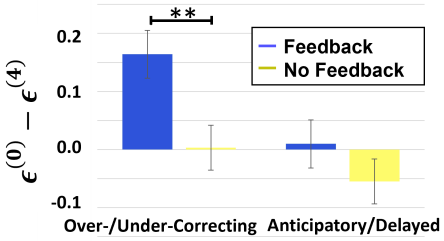


Figure 4.11: Change in $\epsilon_{o/u}$ and $\epsilon_{a/d}$ between first and last calibration tasks.

for each round, i. Figure 4.10 shows a significant correlation ($\rho = .23, p = .002$) between embedding distance and performance, suggesting that the embedding estimated by the EPN is a good measure of suboptimality.

Next, we investigate the overall change in the quality of the participants' demonstrations as measured via the first set of calibration tasks (conducted at the beginning) and last set (conducted at the end). Figure 4.11 shows the change, $\epsilon^{(0)} - \epsilon^{(4)}$, for the o/u dimension and the a/d dimension. We find that participants became significantly better in Feedback ($t(52) = 2.62$, $p = .006$) compared to No Feedback in the o/u dimension. While we do not find significance in a/d, we do find that participants improve in Feedback whereas they become worse in No Feedback in this dimension. Lastly, we investigate Feedback versus No Feedback in terms of subjective metrics. We find that Feedback significantly increases trust ($Z = -2.33$, $p = .019$) and decreases workload [116] ($t(58.0) = -1.79$, $p = .039$) compared to No Feedback. **Takeaway: Feedback derived from our EPN improved participant teaching and resulted in better learning outcomes for the robot in novel tasks.**

4.4.4 Discussion and Limitations

In this work, we aim to determine 1) if we can shift a demonstrator's personalized embedding in one-dimension, 2) how best to provide feedback to shift a demonstrator's embedding in multiple dimensions, and 3) if robotic feedback results in better learning outcomes. In Study

1, we demonstrated we can shift a demonstrator's embedding both farther from and closer to the perfect embedding depending on whether the demonstrator received feedback from an Adversarial robot or a Cooperative robot respectively ($p < .001$). This finding suggests that demonstrators are able to appropriately alter their demonstrations based upon robotic instruction derived from our Reciprocal MIND MELD architecture. Furthermore, the strong correlation between embedding distance and ground truth performance determined via DTW ($p < .001$), suggests that demonstrators are objectively improving their demonstrations in the Cooperative condition and becoming worse demonstrators in the Adversarial Condition.

In Study 2, we explored how best to provide robotic feedback across multiple dimensions of suboptimality to both reduce cognitive overload while providing maximal information. We found that providing feedback intended to improve upon both dimensions of suboptimality simultaneously is the best strategy and does not cause participants to suffer from an undue level of cognitive overload ($p < .001$). Furthermore, participants preferred the Simultaneous robot and rated Simultaneous as better in terms of both trust ($p = .032$) and team fluency ($p = .002$).

Studies 1 and 2 present strong evidence that robotic feedback is capable of improving upon the quality of a teacher's demonstrations, suggesting that a robot will learn better from a teacher who has received robotic feedback. In Study 3, we tested this hypothesis and determine if Simultaneous robotic feedback results in improved learning outcomes for the robot compared to when no feedback is provided. We found that final distance of the robot from the goal improves as the demonstrator receives more feedback about their demonstrations ($p = .045$) whereas we did not find improvement with no feedback. Thus, we have demonstrated that robotic feedback derived from our Reciprocal MIND MELD architecture results in better learning outcomes for a robot.

Limitations: A limitation of Reciprocal MIND MELD is that domain knowledge is required to determine the dimensions of suboptimality. However, robotic domains share many similarities in terms of the control interfaces and the potential for suboptimality,

suggesting that the dimensions in one domain will likely be similar in others. In this work, we investigate verbal feedback to improve upon demonstration quality. However, prior work has suggested that alternative methods of providing feedback may be more effective at improving teaching abilities [60]. We leave to future work an investigation of the best modality for providing demonstrator feedback. Furthermore, we only investigate two dimensions of suboptimality in a driving simulator domain. We leave to future work an investigation of Reciprocal MIND MELD's ability to generalize to additional dimensions of suboptimality in other domains. Additionally, our population consisted mostly of college aged students. In future work, we propose to sample from a more diverse participant pool.

4.5 Conclusion

We introduce Reciprocal MIND MELD, a novel LfD framework for providing robotic feedback to a human demonstrator based upon a personalized embedding to improve suboptimal teaching tendencies. We demonstrate our approach in a series of three human-subject experiments in which we show that robotic feedback can improve upon the quality of a teacher's demonstrations, providing feedback in multiple dimensions simultaneously is the most effective method, and robotic feedback results in improved learning outcomes for a robot. Additionally, we show that our EPN is capable of accurately estimating the updated personalized embedding online, thus enabling continuous feedback to be provided to the demonstrator.

My Reciprocal MIND MELD framework is the first approach to enable personalized teaching to improve the ability of suboptimal demonstrators in an LfD paradigm. Beyond LfD, Reciprocal MIND MELD has the potential to be deployed as a general tutoring and coaching framework to improve end-user suboptimality in a variety of domains.

CHAPTER 5
MANIPULATING AUTONOMOUS VEHICLE EMBEDDING REGION FOR INDIVIDUALS' COMFORT

5.1 Introduction

In Chapter 3 and Chapter 4, I developed personalized frameworks to both learn from heterogeneous, suboptimal demonstrators and also provide robotic feedback to improve upon a teacher's ability to provide high-quality demonstrations. I demonstrated that a personalized framework allows a robot to better learn from a suboptimal demonstrator. The ability to adapt to suboptimal humans and provide tailored feedback to a suboptimal end-user is an important skill for robots to acquire and will enable them to better operate in human-robot teams. However, suboptimality is only one aspect of human heterogeneity which human-machine systems will have to account for. Humans have evolved over thousands of years to have differing preferences for food, habitats and objects and these differing preferences played an integral role in human survival [123]. This innate tendency to have differing preferences, which is known as "evolutionary aesthetics" will affect the way in which humans interact with and regard robots. The way that a robot looks, behaves, and the decisions it makes will all be judged through an individual's unique aesthetic lens. To account for individual preferences, robots should be capable of learning about the unique preferences of the end-user and adapting accordingly [123].

One area in which we see strong differential preferences emerge with regards to robot behavior is in autonomous driving. Prior work has shown that individuals preferences vary greatly with regards to autonomous vehicle driving style [71]. This finding is not surprising considering the diversity that has been demonstrated in individuals' own driving styles. Prior work has broken driving style into numerous categories including aggressive, defensive,

cautious, reckless, anxious, etc [69]. Naturally, such diversity would be reflected in an end-user's preference for AV driving styles.

Driving style is defined as the characteristics of driving related to the judgment and decisions of the driver in a specific situation [68]. Research has shown that driving styles differ greatly amongst individuals [124]. For example, the way in which a driver interacts with other drivers, the level of aggression that a driver exhibits, and tendency to commit traffic violations are characteristics that define an individual's unique driving style. For example, the way in which a driver interacts with other drivers, the level of aggression that a driver exhibits, and the decisions that a driver makes are characteristics that define an individual's unique driving style. Because of these individual differences, when riding in an AV, end-users' expectations and preferences for the behavior of the AV will likely be influenced by their own driving style [125, 126, 29]. One-size-fits-all models employed by AVs which ignore driver differences may lead to decreased acceptance [125]. Instead, the driving style of AVs should be personalized to fit the preferences and expectations of individual end-users.

Prior work has assumed that, to increase end-user acceptance and trust, AVs should mimic end users' unique driving styles [29, 30]. However, even if we are able to personalize an AV's behavior, not all end-users will necessarily want the AV to drive *exactly* as the end-user drives [74, 71]. In fact, prior work has suggested that end-users may want an AV to drive more cautiously than they drive [30, 74, 71]. Additionally, factors such as trust and familiarity with AVs and various personality traits may affect preference for driving styles similar to one's own [127, 128, 129].

From prior work [127, 128, 129, 30], we hypothesize that an individual's optimal driving style is a function of both the end-user's own driving style and various subjective factors. In this work, we develop a data-driven approach capable of producing an optimized driving style for an end-user based upon their driving style and relevant subjective factors.

We introduce MAVERIC, a learning from demonstration approach for personalizing the

Figure 5.1: 6-DOF driving simulator developed by Toyota Research Institute.

driving style of an AV. By observing the driving of an end-user, MAVERIC learns a high-level model via a neural network architecture that predicts personalized control parameters for low-level controllers. Simultaneously, MAVERIC learns a personalized embedding representing the driving style of an end-user. By shifting the personalized embedding along the gradient of aggression, MAVERIC tunes the AV driving style to be more aggressive or cautious while maintaining other personalized characteristics. This capability allows us to modulate the AV's aggressive driving style with respect to an end-user's style so as to optimize the AV's driving style.

In two human-subjects studies, we investigate if MAVERIC can effectively mimic an individual's driving style as well as modulate aggression. Additionally, we investigate the factors that influence the effect of *homophily* - i.e., preference for a driving style similar to one's own. We demonstrate that preferred driving style is related both to one's own style as well as personality traits, perceived similarity, and high-velocity driving.

In this work we contribute the following:

1. We formulate MAVERIC, a novel framework to personalize driving style and modulate aggressiveness while maintaining other aspects of driving style.

2. We demonstrate that MAVERIC can closely match an end-user's driving style ($p < .001$) as well as produce more aggressive ($p < .001$) and more cautious ($p < .001$) driving in a high-fidelity driving simulator.

3. We find that personality ($p < .001$), perceived similarity ($p < .001$), and high-velocity

driving style ($p = .0031$) significantly impact the effect of homophily.

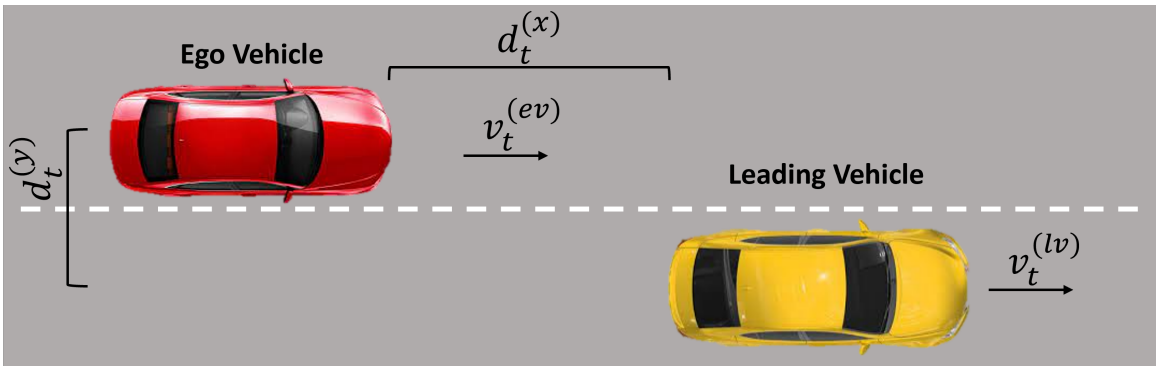


Figure 5.2: This figure shows our domain of light traffic and associated state information. $v_t^{(ev)}$ is the velocity of the ego at time t, $v_t^{(lv)}$ the velocity of the leading vehicle, $d_t^{(x)}$ the distance between the leading vehicle and ego in the x direction at time t, and $d_t^{(y)}$ the distance in y at time t.

5.2 Methodology

In the following section we provide an overview of MAVERIC. We discuss our architecture and how we endow our framework with the ability to modulate aggression. Figure 5.2 depicts the state information relevant to our architecture.

5.2.1 Network Architecture

Our network architecture is depicted in Figure 5.3. Our network simultaneously learns the high-level parameters of low level controllers and the personalized embedding, $w^{(p)}$, representing the driving style of an individual, p. See Section Subsection 5.2.3 for details on the low-level controllers. Our network is composed of five subnetworks: the Lane Change Predictor, C_ψ, Velocity Predictor, V_β, Following Distance Predictor, F_ϕ, Style Predictor, S_θ, and Mutual Information, M_α. All subnetworks consist of linear layers with ReLU activations and are trained via a mean-squared error loss unless otherwise specified.

Lane Change, Velocity, and Following Distance Predictors: The Lane Change (C_ψ), Velocity (V_β), and Following Distance (F_ϕ) Predictor subnetworks each take as input the

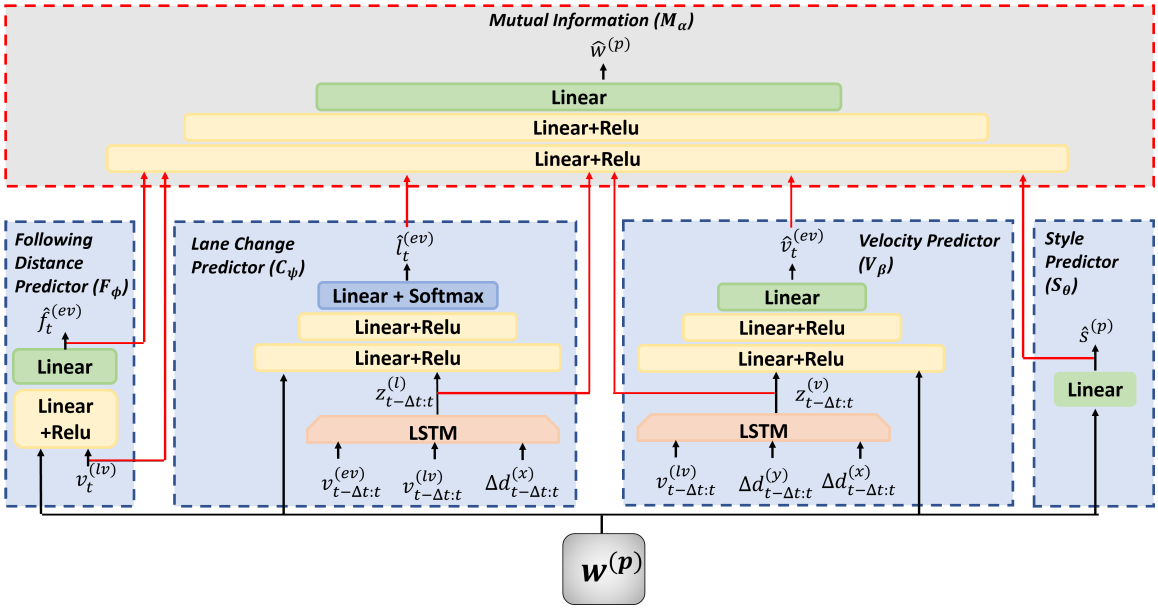


Figure 5.3: This figure shows our network architecture. F_ϕ predicts the following distance. C_ψ predicts when a lane change should occur for the ego vehicle. V_β outputs the velocity of the ego vehicle. S_θ is the style predictor subnetwork which predicts the subjective aggressive style of the participant from the personalized embedding, $w^{(p)}$. $v^{(ev)}_{t-\Delta t:t}$ is the ego velocity and $v^{(lv)}_{t-\Delta t:t}$ is the velocity of the lead vehicle from time $t-\Delta t$ to t. $d^{(x)}_{t-\Delta t:t}$ is the distance between the ego and leading vehicle in x and $d^{(y)}_{t-\Delta t:t}$ is the distance in y. $\hat{w}^{(p)}$ is the estimate of the participant's personalized embedding sampled from the approximate posterior defined by M_α.

personalized embedding, $w^{(p)}$, and the relevant states as shown in Figure 5.3. C_ψ and V_β utilize a LSTM network to predict the probability, $\hat{l}_t^{(ev)}$, of a lane change occurring and the desired velocity of the ego vehicle, $\hat{v}_t^{(ev)}$, respectively. We utilize a Softmax activation for the last layer of C_ψ and this subnetwork is trained via a cross-entropy loss. F_ϕ predicts the desired following distance, $\hat{f}_t^{(ev)}$, between the ego vehicle and the lead vehicle.

Style Predictor: The Style Predictor subnetwork, S_θ, takes as input the personalized embedding, $w^{(p)}$, and is trained to predict the subjective aggressiveness, $\hat{s}^{(p)}$, of the participant, p. In Section Subsection 5.2.2, we discuss the importance of this subnetwork and how we obtain $s^{(p)}$.

Mutual Information: The Mutual Information subnetwork, M_α, seeks to maximize mutual information between the driving style of the individual and $w^{(p)}$ so as to ensure

that the learned embedding captures the differences in driving style between individuals [104, 1, 2]. M_α takes as input encodings $z^{(l)}_{t-\Delta t:t}$ and $z^{(v)}_{t-\Delta t:t}$ and the outputs of each of the other subnetworks, $\hat{l}_t^{(ev)}$, $\hat{v}_t^{(ev)}$, $\hat{f}_t^{(ev)}$, and $\hat{s}^{(p)}$. The subnetwork learns to map these inputs to $\hat{w}^{(p)} \sim \mathcal{N}(\mu^{(p)}, \sigma^{(p)})$.

5.2.2 Modulating Aggression

We designed MAVERIC to be capable of both matching driving styles of individuals and modulating aggression with respect to an individual's driving style. Because MAVERIC learns a latent embedding space, we can create a dimension of aggression within the embedding space, allowing us to shift an embedding along that dimension and modulate aggression, while keeping other driving characteristics constant. To achieve this, we add an additional signal when learning the embedding space. We add a network head, S_θ, composed of a linear layer which takes as input the personalized embedding, $w^{(p)}$. S_θ is trained to predict the subjective aggressive driving style of the end-user as measured by the participant's response to the Aggressive Driving Behavior (ADB) scale [73]. We train the network to predict the aggressive driving style, thereby creating an aggressive dimension within the embedding space. We can then move along the gradient of aggression (∇S_θ) to produce a more or less aggressive driving style as shown in Figure 5.4.

While driving style has multiple dimensions [69], we focus on the aggressive dimension, as prior work has shown that this dimension has a large impact on end-user preference [71, 30, 74]. Other characteristics of driving could be modulated by following a similar procedure. While we acknowledge that the ADB scale is a noisy metric, as discussed in Section 5.4, our results demonstrate that our method can effectively produce more and less aggressive behavior.

5.2.3 Low Level Controllers

MAVERIC learns the parameters for low-level controllers (e.g., velocity, timing of lane change, etc.) rather than directly learning the low-level control inputs (i.e., throttle and steering) to enable safety constraints and account for unexpected or dangerous behavior that could be produced by the network. For example, by learning the desired following distance for an end-user and utilizing an adaptive cruise controller to maintain this distance, we can ensure that the following distance remains safe. Additionally, by predicting when a lane change should occur via the neural network and utilizing a low-level controller to execute the lane change, we ensure consistent and smooth lane changes. Furthermore, this hierarchical method of learning and control has been shown to produce better results in prior work [130].

Lane Change Controller: Our lane change controller is based on a Stanley controller [107] and follows a Bezier curve [131]. We compute the Bezier curve based upon the desired distance (selected to produce natural behavior) to complete the lane change, while ensuring that the ego vehicle will not collide with the leading vehicle. The lane change controller executes a lane change when $\hat{l}_t^{(ev)} > \delta$.

Velocity Controller: We utilize a Proportional and Integral (PI) controller to maintain the desired velocity, $\hat{v}_t^{(ev)}$ of the ego vehicle as predicted by the neural network.

Following Distance Controller: When the distance between the ego and leading vehicle falls below threshold, λ, we switch from the velocity controller to the following distance controller. The following distance controller is a PI controller that minimizes both the error between the desired following distance, $\hat{f}_t^{(ev)}$, as predicted by the neural network and the difference in speed between the ego and leading vehicle subject to safety constraints on following distance.

5.3 Human Subjects Studies

We conducted two human subjects studies: A Model Training Study (Study 1) and a Model Testing Study (Study 2). In Study 1, we collect data from 30 participants to train MAVERIC and learn θ, ϕ, ψ, α, and β, and the participants' embeddings. In Study 2, we collect driving data from 24 participants to learn their embeddings and then each participant experiences the four AV conditions (Subsection 5.3.4). Additional details for our studies can be found in [132].

5.3.1 Driving Simulator

To test the abilities of MAVERIC, we utilize a high-fidelity driving simulator developed by TRI.[1] The simulator (Figure 5.1) is an immersive 6-DOF platform capable of emulating the motion of a vehicle. The simulator is based on CARLA [133], ROS2, and Unreal Engine.

5.3.2 Participants

Model Training Study (Study 1): We recruited 30 participants (Mean age 35.4; 27% Female) from TRI and Woven Planet via word of mouth and mailing lists. Four of the participants were professional drivers who demonstrated aggressive, cautious, and their own driving style. In total, we collected 38 data points representing various driving styles.

Model Testing Study (Study 2): Study 2 was run with two different populations of participants to increase diversity. For the *internal* study, 12 subjects were recruited from TRI and Woven Planet (Mean age 34.42; 33.4% Female). For the *external* study, 12 subjects (Mean age 43.92; 41.7% Female) were recruited from the general public via Fieldwork recruiting. Except where noted, the studies are analyzed as a collapsed dataset because the procedure was identical.

[1] https://medium.com/toyotaresearch/driver-in-the-loop-simulation-for-guardian-and-chauffeur-847f36ea103e

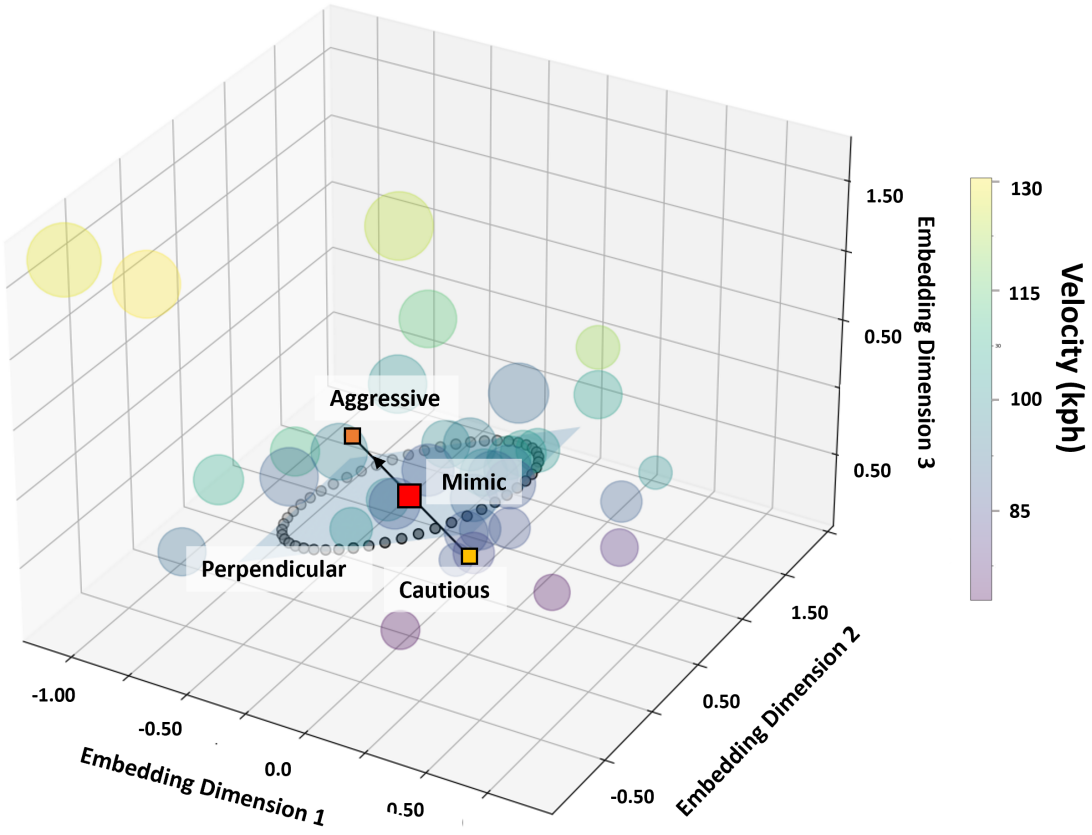


Figure 5.4: This figure shows the learned embedding space. The size of the points represents the subjective aggressive style of the participant and color represents the average velocity. The black line shows the vector of the aggressive gradient. The red square represents a candidate learned embedding of a participant. We shift the embedding along the gradient to increase (orange square) or decrease (yellow square) the ADB score by 15 points to produce behavior for the aggressive and cautious conditions respectively. We randomly sample from the gray points to produce the Perpendicular behavior.

5.3.3 Procedure

We investigate personalization of driving styles in the domain of light traffic on a two-lane highway (Figure 5.2). Research was approved by WCG IRB (protocol #20221727). External participants were compensated $250.

Model Training Study (Study 1): Participants control the vehicle and demonstrate their driving style for 10 minutes. Their task is to drive as they would in their own vehicle. They are instructed to maintain the speed they would typically drive if the speed limit is 55mph

and to pass other vehicles when they feel it is appropriate. In this domain, participants encounter vehicles in the same lane (leading vehicles) and in the adjacent lane (off-lane vehicles). Participants must make decisions about changing lanes, following distance, and velocity. The speed of the leading vehicles is randomly selected without replacement from the set $\{0.85v_e, 0.9v_e, 0.97v_e, 0.9s, s, 1.1s\}$ where v_e is the ego target speed and s the posted speed (55mph). These speeds ensure consistency across participants but also ensure that some of the leading vehicles are slower than the ego, thus forcing the participant to make a decision about changing lanes.

Participants first complete pre-study surveys to collect information about demographics and attitude towards AVs (Subsection 5.3.5). Participants complete a practice session to familiarize themselves with the vehicle controls and domain. We next collect driving data from the participants to learn the network parameters and their personalized embeddings.

Model Testing Study (Study 2): In the testing study, we freeze the network parameters, ϕ, θ, ψ, β, and α learned from Study 1 data. Participants fill out the pre-study surveys and then drive the vehicle in the highway domain. We collect their data to learn their embedding. This procedure is the same procedure experienced by training participants. Participants next experience four AV conditions as described in Subsection 5.3.4. After each condition, participants fill out surveys about their subjective perception of the AV (Subsection 5.3.5).

5.3.4 Model Testing Study Conditions

The behaviors described below are created by shifting a participant's embedding in the embedding space. Figure 5.4 shows the learned embedding space and how we choose the embedding to create the behavior for each of the conditions. We hypothesize that Mimic will produce similar behavior relative to the participant's driving, Aggressive will produce more aggressive behavior and Cautious, less aggressive.

Mimic: In *Mimic*, we utilize the personalized embedding learned from the participant's data to produce driving behavior to mimic the participant's own driving style.

Aggressive: In *Aggressive*, we shift the participant's embedding in the positive gradient of aggression (equivalent to fifteen points on the ADB survey) to produce more aggressive behavior while maintaining other characteristics of driving style (i.e., $S_\theta(\hat{w}^{(p)}) = \hat{s}^{(p)} + 15$). We constrain $\hat{s}^{(p)} + 15$ to be no more than the largest possible score on the ADB survey (55 points).

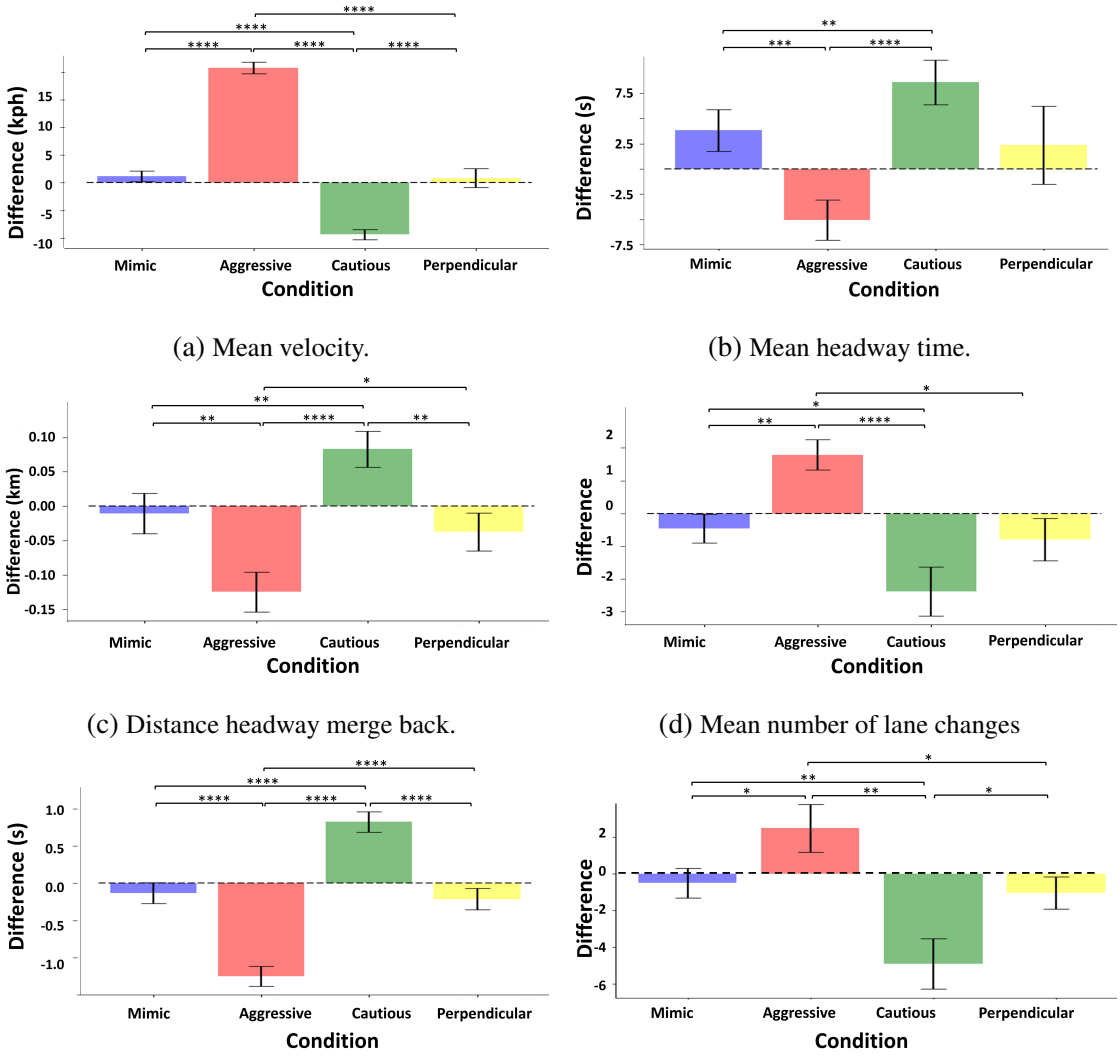
Cautious: In *Cautious*, we shift the embedding in the negative gradient of aggression ($S_\theta(\hat{w}^{(p)}) = \hat{s}^{(p)} - 15$) to produce less aggressive behavior while maintaining other characteristics of style. We constrain $\hat{s}^{(p)} - 15$ to be no less than the smallest possible score on the ADB (11 points).

Perpendicular: We include *Perpendicular* to conduct an exploratory investigation into the behavior produced when we maintain the level of aggression but move the embedding within the plane perpendicular to the aggressive gradient. Our objective is to investigate which driving characteristics change as a result of this shift. To select the embedding, we randomly sample a point along an ellipse on the plane one standard deviation away from the participant's embedding as shown by the gray points in Figure 5.4. By doing so, we are able to keep the degree of aggression constant, while altering other aspects of driver style. We hypothesize that Perpendicular will produce similarly aggressive behavior compared to the participant's driving.

5.3.5 Metrics

Participants in both Study 1 and Study 2 complete the pre-study surveys. Only participants in Study 2 complete the post-trial surveys. The surveys detailed below comply with the design guidelines outlined in Schrum et al. [113] and are validated from prior work when possible.

Pre-study: The pre-study survey is intended to measure the participants' subjective attitudes towards AVs. We collect demographic information and Big-Five personality information via the Mini International Personality Item Pool [134]. To measure a participant's



(a) Mean velocity.

(b) Mean headway time.

(c) Distance headway merge back.

(d) Mean number of lane changes

(e) Time headway merge back

(f) Subjective aggressive rating

Figure 5.5: This figure depicts the difference between the AV's driving style and the participants' driving style for our objective and subjective metrics. We show that both objectively and subjectively our approach can mimic an individual's driving style as well as modulate aggression.

aggressive driving style, we utilize the Aggressive Driving Behavior Scale [73]. We measure other aspects of driving style via the Multi-Dimensional Driving Style Inventory [69] and measure experience with cars/racing games/AVs [2], trust in AVs [135], perception of AVs [136], and trust in automation [137].

Post-trial: The post-trial surveys capture the participants' subjective attitudes towards each of the AV conditions. We measure perceived intelligence [138], competence [139],

discomfort [139], and trust [135]. We modify each of the subscales for AVs. Additionally, we create two custom scales to measure perceived similarity and aggressiveness relative to the participant's own driving style.

Objective Measures: In keeping with prior work [71, 140], we measure various metrics to determine how similar the driving style of each condition is compared to the participant's own driving style. We investigate mean velocity and mean number of lane changes. We also measure mean headway time (the distance between the lead vehicle and ego divided by the speed of the ego when a lane change occurs), distance headway merge back (the distance between the following vehicle and ego), and time headway merge back (distance headway merge back divided by the speed when a lane change occurs).

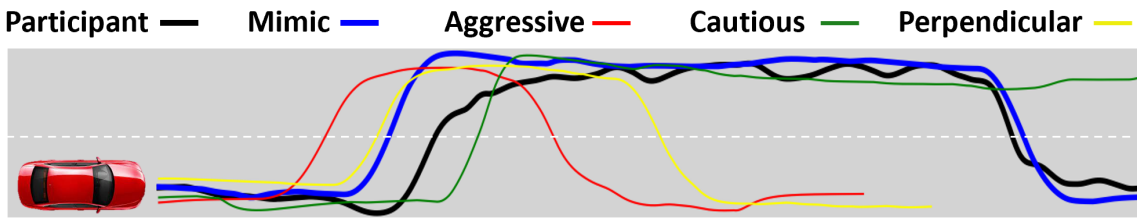


Figure 5.6: This figure shows the trajectories generated by the four conditions compared to the participant's demonstration (black).

5.4 Results

Figure 5.6 shows an example of the trajectories produced by the four conditions. Mimic changes lanes to pass the leading vehicle at a similar time as the participant whereas Aggressive and Cautious change lanes sooner and later respectively. Because of Cautious's lower speed, the AV never passes the leading vehicle. Perpendicular executes a lane change at a similar time to that of the participant. However, the AV changes back to the right lane sooner.

5.4.1 Analysis of Embedding Space and Aggressive Gradient

We first investigate if our embedding space is capable of representing and producing diverse driving styles and if the aggressive gradient correlates with relevant objective metrics. To investigate these questions, we project the learned embeddings of the test participants onto the line representing the gradient of aggression. We then analyze how driving style changes as a result of the position of the embedding along this line. We find that as we move along the aggressive gradient, the average velocity of the participant increases. The average velocity along the aggressive gradient ranges from 54.5 mph (in the most negative direction of the gradient) to 78.56 mph (in the most positive direction of the aggressive gradient. We find a strong correlation ($r = .49, p = .022$) between the embedding's position along the aggressive gradient and the average velocity of the participant. This finding suggests that, in keeping with prior work [141], velocity is an important component of aggression within the embedding space. We find similar results for mean headway time ($r = -.46, p = .032$), distance headway merge back ($r = -.43, p = .046$), mean number of lane changes ($r = .47, p = .028$), and time headway merge back ($r = -.48, p = .025$). Lastly, we show that a participant's subjective aggressive rating of their own driving style strongly correlates with the position of their learned embedding along the aggressive gradient ($r = .92, p < .001$). These findings provide evidence that our embedding space is capable of representing diverse driving styles and that aggressiveness objectively and subjectively increases as we move along the aggressive gradient.

5.4.2 Algorithm Validation

We next investigate MAVERIC's ability to mimic end users' driving styles and produce more and less aggressive behavior in terms of both objective and subjective metrics. In our following analysis, we verify that data complies with assumptions before applying a parametric test. We first investigate MAVERIC's ability to accurately mimic driving style. We find that the accuracy with which we are able to mimic the participant's velocity is

93.6%, time headway is 80.2%, distance headway merge back is 92.4%, mean number of lane changes is 81.0%, and time headway merge back is 81.8%.

Figure 5.5 shows the differences in our objective and subjective metrics between the participant's driving style and the behavior produced by our four conditions. To determine if there are significant differences between conditions for each of the metrics, we conduct a repeated measures ANOVA with Holm's post hoc correction or a Friedman's test when the data fails assumptions. We find that the difference between Mimic and the participant's driving is significantly less compared to Aggressive and Cautious for all objective metrics ($p < .001$) (Fig Figure 5.5(a) - Figure 5.5(e)). We find that Aggressive maintains a higher velocity compared to Mimic. Additionally, as predicted by prior work [71, 140], aggressive achieves a lower headway merge back time and headway merge back distance. Furthermore, Aggressive commits more lane changes compared to Mimic despite encountering the same number of leading vehicles. We find opposite results with the Cautious condition. We illustrate that the characteristics of our AV driving styles align with the characteristics indicative of aggression in prior work, suggesting that our approach can effectively modulate aggression with respect to one's own driving style [71, 140, 141].

Additionally, as shown in Figure 5.5(f), we find that participants rate Cautious as significantly less aggressive compared to Mimic ($p = .002$) and Aggressive as significantly more aggressive ($p = .017$). Furthermore, we find that Mimic and Perpendicular are rated as similarly aggressive compared to the participant's own driving. **Our objective and subjective results together support our hypotheses that 1) our approach is capable of mimicking driving style and 2), by shifting a participant's learned embedding along the aggressive dimension, we are able to produce objectively and subjectively more aggressive and cautious behavior.**

5.4.3 Maintaining Other Aspects of Driving Style

One of the goals of our approach is to modulate aggression while maintaining other aspects

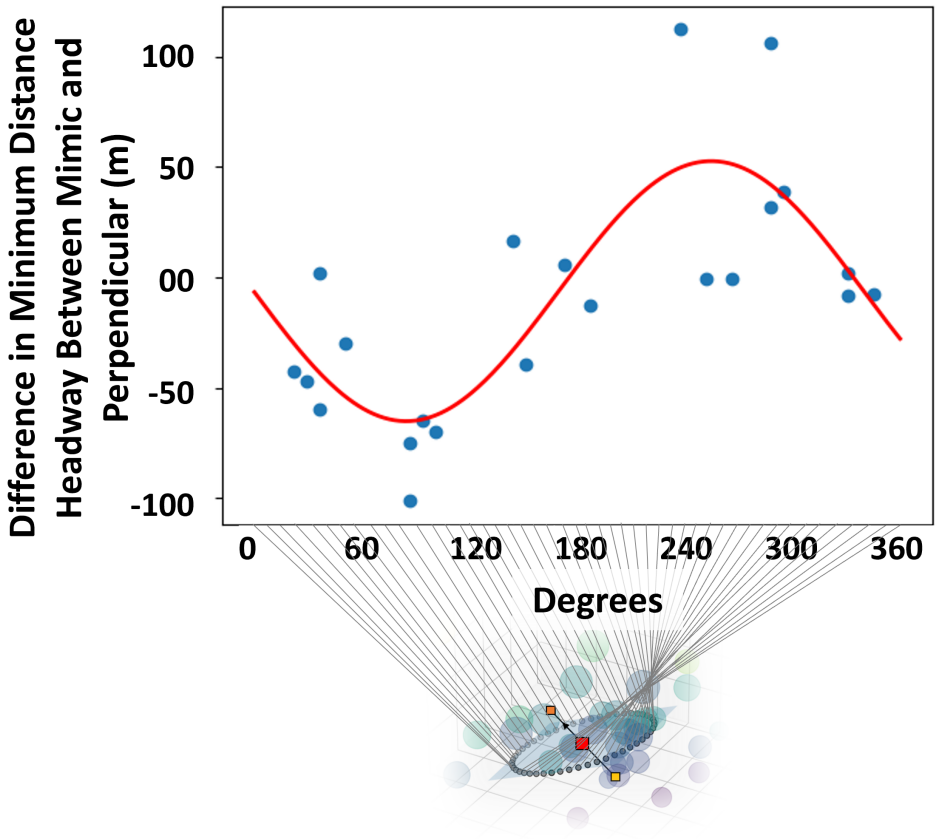


Figure 5.7: This figure shows the changes in minimum headway distance as we move around the ellipse within the plane perpendicular to aggression. Minimum headway distance was not significantly correlated with aggression (Subsection 5.4.3) and is modulated by moving in the plane perpendicular to aggression.

of driving style. If moving along the gradient of aggression modulates the aggressive aspect of the driving style, then we hypothesize that moving within the plane perpendicular to aggression will modulate other aspects of driving style unrelated to aggression. Interestingly, we found that minimum headway distance and fraction of time in the left lane were not significantly correlated with the embeddings position along the aggressive gradient. Moving along the gradient does not significantly alter minimum headway distance or fraction of time in the left lane, suggesting that, in our learned embedding space, these factors do not play a large role in aggressiveness. Therefore, we predict that these aspects of driving will instead be modulated when we move perpendicular to the gradient of aggression. To test this

Table 5.1: This table shows our correlation analysis. M represents Mimic, A represents Aggressive, and C represents Cautious.

Independent	Dependent	Statistic	p-value
Conscientious	M-C Competence	$\rho(22) = -.71$	$p < .001$
Conscientious	M-C Intelligence	$\rho(22) = -.5$	$p = .012$
Conscientious	M-C Discomfort	$r(22) = .46$	$p = .024$
Conscientious	M-C Trust	$\rho(22) = -.62$	$p = .0011$
Conscientious	M-A Competence	$\rho(22) = -.51$	$p = .011$
Conscientious	M-A Intelligence	$\rho(22) = -.51$	$p = .01$
Conscientious	M-A Discomfort	$r(22) = .45$	$p = .028$
Conscientious	M-A Trust	$\rho(22) = -.48$	$p = .0018$
Openness	M-A Discomfort	$\rho(22) = .49$	$p = .015$
Similarity	Trust	$\rho(94) = .16$	$p = .001$
Similarity	Intelligence	$\rho(94) = .34$	$p < .001$
Similarity	Competence	$\rho(94) = .27$	$p < .001$
High-Velocity	M-A Intelligence	$\rho(94) = -.58$	$p = .0031$
High-Velocity	M-A Competence	$r(22) = -.44$	$p = .03$
High-Velocity	M-A Trust	$r(22) = -.43$	$p = .036$

hypothesis, in Fig Figure 5.7 we plot the difference in minimum headway distance between Mimic and Perpendicular versus the position around the ellipse that is depicted in Figure 5.4. We find that minimum headway distance does in fact correlate with position around the ellipse ($r = .68, p < .001$). We additionally find that the fraction of time in the left lane significantly correlates with position around the ellipse ($r = -.47, p = .025$).

We note that minimum headway distance is often associated with aggression [142]. However, this is most often the case when the ego vehicle is not capable of changing lanes and is instead forced to following a leading vehicle. We hypothesize in our work that minimum headway distance is not correlated with aggression because the participant can choose to change lanes at any point to pass a slower driver and therefore is not forced to maintain a following distance if they do not want to.

5.4.4 Homophily

As shown in Figure 5.8 not all participants preferred the Mimic condition. More than 20% of participants preferred the Aggressive condition and more than 25% of participants preferred

the Cautious condition. To explain this finding, we next explore the factors that modulate the effect of homophily (Table 5.1) to determine why some participants prefer a driving style different from their own. First we investigate if a participant's personality impacts their preference via a correlation analysis. As shown in Table 5.1, we find a strong correlation between conscientiousness (i.e., the extent to which one is responsible and dependable [143]) and the difference between a participant's perceived competence of Mimic compared to Cautious, suggesting that individuals higher in conscientiousness prefer a more cautious style to their own. This finding may explain why 62.5% of participants rated Mimic as less than or equal in competence relative to Cautious. To further support the hypothesis that conscientiousness influences the effect of homophily, we find that participants who are higher in conscientiousness rate a more cautious style as significantly more intelligent, comfortable, and trustworthy compared to Mimic and significantly more competent, intelligent, comfortable, and trustworthy compared to Aggressive.

We additionally find that openness (the degree to which one is broad-minded [143]) correlates with the difference between a participant's comfort with Aggressive compared to Mimic. This finding suggests that those who are more open to new experiences may prefer a more aggressive AV and may explain why 37.5% of participants rated Aggressive as causing greater comfort compared to Mimic.

Prior work suggests that *perceived similarity* to one's own driving style is an important aspect of AV acceptance [71]. To investigate this claim, we conduct a correlation analysis between perceived similarity and an end-users preference for the AV. We find a positive correlation between perceived similarity and trust, intelligence, and competence. This finding suggests that perceived similarity should be taken into consideration when optimizing AV driving style.

Prior work demonstrated that one's own driving style may impact preference for an AV's style (e.g., more aggressive drivers prefer relatively less aggressive AVs) [71, 74]. To investigate this question further, we conduct a correlation analysis between the dimensions

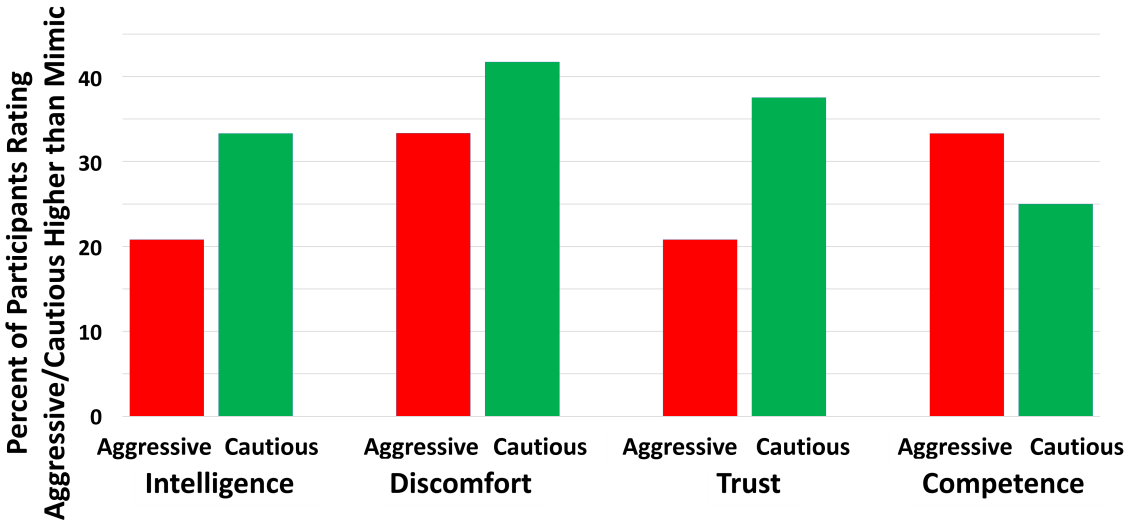


Figure 5.8: This figure shows the percent of participants who rated Aggressive and Cautious as better than Mimic in terms of each of our subjective metrics.

of the Multi-Dimensional Driving Style Inventory [69] and preference for Aggressive and Cautious compared to Mimic. We find that participants who report a high-velocity driving style rate Aggressive to be significantly higher than Mimic in intelligence, competence, and trustworthiness. This findings suggests that high-velocity drivers may prefer a more aggressive AV. Due to the contradictory findings with prior work, we aim to conduct a deeper analysis into how the specific dimensions of one's own aggressive style impact the effect of homophily in future work.

We note that the results we present are an exploratory analysis and we do not claim to demonstrate a causal relationship between the subjective factors discussed above and homophily. However, our findings suggest that these factors warrant further investigation in future work. **Overall, our findings demonstrate that personality traits, perceived similarity, and high-velocity driving style may be important factors in modulating the effect of homophily.**

5.5 Discussion

Our results demonstrate the our MAVERIC framework is capable of both mimicking and

modulating driving style by learning an embedding representing an end-user's own driving style. Given other relevant factors related to end-user characteristics, we can then tune this driving style to better match the preference of the end-user. Thus, while other approaches either directly mimic the end-user's own driving style or do not take into consideration the end-user's driving style at all, our approach is capable of integrating both information about an end-user's own driving style and subjective characteristics that are predictive of the optimal AV driving style.

In our analysis, we show that our learned embedding space captures salient aspects of driving style and that the gradient of aggression correlates with objective and subjective aggressive metrics. An interesting aspect of our aggressive dimension is that this representation of aggression is not based on a pre-defined or hand-crafted heuristic but is instead based upon end-users' perception of what is meant by aggressive driving. By defining aggression via this subjective metric, we are able to produce driving styles that are perceived to be more aggressive or more cautious by end-users.

In our analysis of the effect of homophily, we aim to determine the subjective factors that future work should consider when optimizing driving style. We show that simply mimicking an end-user's own driving style is often not preferred and that certain subjective characteristics may explain the discrepancy between an end-user's own driving style and their preferred AV driving style. By conducting a correlation analysis, we uncover several characteristics that impact homophily. We find that personality should be considered when determining the optimal driving style and that specifically, conscientiousness and openness to experience are important factors. Additionally, participant's perception of their own driving style, e.g. self-reported high-velocity driving style may influence an individual's preference for a more aggressive driving style. We additionally find that perceived similarity is a relevant factor as supported in prior work [71]. These findings provide us with insight into what factors should be considered when determining exactly how much and in which direction to shift an end-user's personalized embedding along the aggressive gradient so as

to optimize driving style.

5.6 Limitations and Future Work

In future work we aim to quantify the relationship between relevant subjective factors and the preferred level of aggression. By doing so, we will be able to determine exactly how much to shift an individual's embedding along the aggressive dimension so as to produce the optimal driving style for an individual. Additionally, we plan to investigate MAVERIC's abilities to learn driving styles in domains involving more traffic and the potential for more complex decision making. A limitation of our work is that we only recruited internal participants for Study 1. However, despite this limitation, our study comprises a more diverse population pool than many studies in human-robot interaction which typically recruit from a pool of college students [144]. Another limitation is that the perceived similarity and aggressiveness surveys are not verified in prior work. Additionally, because we only conduct a correlation analysis, we cannot conclude that the subjective factors are causally related to homophily. However, our results suggest that these factors are worthy of further investigation in future work.

5.7 Conclusion

We have presented MAVERIC, a novel framework to personalize driving style and modulate aggressiveness. We demonstrated MAVERIC's ability to reproduce an end-user's own driving style and investigated how the preference for one's own style is modulated by personality, perceived similarity, and high-velocity driving style. To our knowledge, ours is the first framework to combine subjective metrics with end-user training data to produce a personalized AV controller. Our results indicate that personalizing AV control is a research area that merits further investigation and may provide a path towards greater AV acceptance.

My MAVERIC framework fills a gap in prior work by enabling personalization of AV driving styles beyond simple mimicry. By taking into account relevant end-user charac-

teristics to determine the optimal level of aggression, MAVERIC is capable of optimizing driving style for an end-user and thereby improve end-user experience and acceptance.

CHAPTER 6

SAFE META ACTIVE LEARNING FOR DEEP BRAIN STIMULATION

6.1 Introduction

In his book *I, Robot*, Isaac Asimov states that the First Law of Robotics is that "a robot may not injure a human being or, through inaction, allow a human being to come to harm" [145]. This law, though developed for a fictional setting, must be taken seriously by robotics researchers to ensure safety of the human end-user [146]. Thus far in this thesis, I have developed novel algorithms to personalize human-machine interaction. Yet we have not explicitly considered safety critical domains in which machines must not only adapt to humans, but must do so safely.

Personalization of robots and autonomous systems in the domain of healthcare is critical. Patients experience different disease manifestations as well as differing needs based on biology, age, and other factors and therefore, human-machine systems will have to adapt to meet the requirements of individual patients. However, personalization is not the only important consideration in healthcare. The safety of patients is paramount when deploying machines to support patient care. Therefore, these systems must be equipped with strong safety guarantees to ensure the well-being and safety of the patient.

Deep Brain Stimulation (DBS) is an example of a healthcare application which requires the machine to both adapt to the patient's biology as well as ensure patient safety. DBS devices implanted in the brain can improve memory deficits in patients with Alzheimers [147] and responsive neurostimulators can counter epileptiform activity to mitigate seizures. However, because the brain anatomy, electrode placement and other variables differ across patients, the relationship between the electrode parameters and clinical outcome is not consistent across patients. Therefore, surgeons must manually test various stimulation pa-

rameters for each patient, a process which is time-consuming and laborious, often resulting in extended patient suffering. Furthermore, certain parameter settings can be risky and possibly trigger an ictal state or brain damage. The ability to efficiently and safely learn a model of the brain which maps parameter setting to seizure reduction has the potential to improve clinical outcomes and the lives of patients.

To address the problem of sample efficient learning, researchers have previously investigated active learning techniques for [148, 149]. However, prior work in active learning suffers from three weaknesses: 1) an inability to accurately quantify expected informativeness [94], 2) a lack of generalizability [97], and 3) a lack of safety considerations [33]. Active learning approaches typically hand-engineer heuristics or acquisition functions to select the best action [32, 33]. However, these heuristics are only proxies for true informativeness of a data point and may not accurately quantify the actual informativeness of a data point when updating the model with this new training data. Additionally, heuristics that are well suited for one active learning domain may not be effective in another. The few meta-active learning approaches proposed in recent years rely only on hand-engineered features which reduce generalizability and require expert feature selection [94]. Furthermore, prior approaches do not consider applications in safety critical domains in which constraints must be placed on the acquisition function to prevent the model from sampling unsafe configurations [150].

Yet, efficient learning is not the only criteria that must be met when dealing with safety critical domains. For example, when learning the optimal parameter settings for a DBS pateint, one must reason about the safety of the patient in addition to expected informativeness of an action to prevent the patient from experiencing unwanted or dangerous side-effects. If the optimal parameter settings can be learned efficiently and safely, the patient may experience improved symptoms without negative side-effects.

To achieve the goal of safe and efficient adaptation, in this chapter, I introduce Safe MetAL, a hybrid meta-learning and mathematical programming approach that enables efficient, safe, and computationally fast optimization of a latent human-machine system.

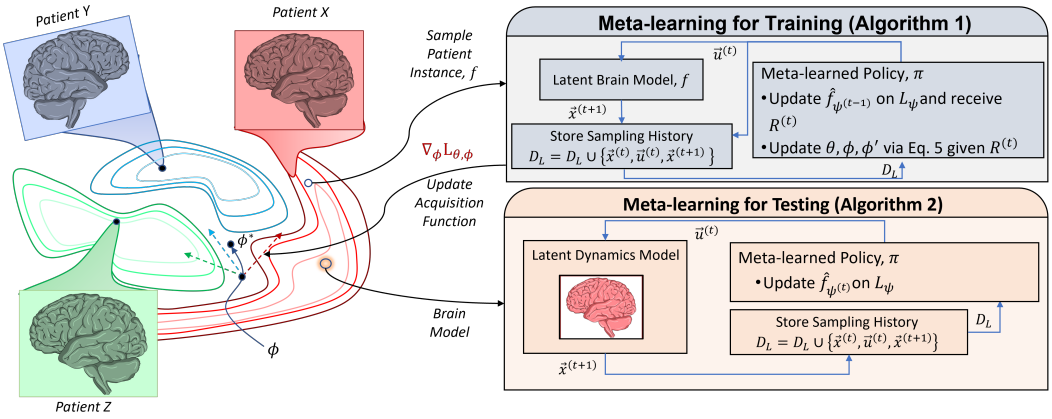


Figure 6.1: This figure shows our meta-learning framework, grounded in our DBS application. The red, blue and green curves represent the hypothetical manifolds within our distribution of patients. Our meta-learning algorithm samples from the distribution of patients and learns a function, Q_ϕ, describing the expected informativeness of taking an action. We embed this Q_ϕ in a MILP to enforce safety-constraints, thereby ensuring patient safety while enabling efficient learning of the optimal parameters.

The key to our approach is that we safely and efficiently meta-learn an acquisition function based on a learned representation of sample history that accurately quantifies the expected informativeness for an unknown, latent model when taking a given action and experiencing the resultant state. By directly encoding this acquisition function in a chance-constrained mixed-integer linear program (Mixed-Integer Linear Program (MILP)), we can simultaneously enforce safety guarantees [33] while taking an action which maximizes expected informativeness. This acquisition function is meta-learned offline over a distribution of tasks, allowing the policy to benefit from past experience and provide a more robust measure of the value of a labeled data point.

We demonstrate the advantage of Safe MetAL for deep brain stimulation, a safety critical environment in which personalization, safety, and sample efficiency are of utmost importance. Our approach outperforms previous Bayesian [18, 151], meta-learning [94, 151], and active learning approaches [32, 33] in terms of expected informativeness, safety, and computation time.

Contributions:

1. We present Safe MetAL, a meta-learning algorithm for learning the personalized parameter settings for DBS patients via a domain-specific acquisition function that accurately quantifies expected informativeness. Safe MetAL (1) meta-learns an acquisition function to quantify domain specific expected informativeness of a data point without the need for hand-derived features and ad hoc engineering, and (2) reasons explicitly about exploitation vs. exploration by trading off gaining information and probabilistically-safe control.
2. We formulate a novel bridge between deep learning and mathematical programming techniques in a way that is fully, end-to-end differentiable and trainable by embedding this meta-learned acquisition function within a chance-constrained optimization framework to achieve probabilistic guarantees.
3. We show that our approach sets a new state-of-the-art for model accuracy (41%) compared to Bayesian [18, 151], active [32, 33] and meta-learning [94, 151] approaches and computational speed ($+20\%$) versus two active and meta-learning baselines while also providing probabilistic guarantees.

6.2 Problem Set-up

We describe our problem set-up in the context of DBS. Our objective is to safely and efficiently learn the model of the patient's brain (i.e., the mapping from parameter setting to brain state), $\hat{f}_\psi$, and thereby determine the optimal parameter settings. Specifically, we seek to determine the parameter setting that should be applied next to provide maximum information about the nature of the parameter-brain state relationship given the patient's previously experienced brain states and parameters without causing unwanted side-effects. To do so, we learn a function that describes the expected informativeness when applying any parameter setting, conditioned on the prior parameters applied to the patient and subject to safety considerations. We set up our problem in three steps: (1) active learning, (2) safety, and (3) meta-learning.

First, we define our unlabeled dataset, $D_U = \langle \vec{s}^{(i)}, \vec{a}^{(i)} \rangle_{i=1}^{n}$, as consisting of all possible state-parameter pairs that the patient could potentially experience and the labeled dataset, $D_L = \langle \vec{s}^{(i)}, \vec{a}^{(i)}, \vec{s}^{(i+1)} \rangle_{i=1}^{m}$, as the set of state transition triples previously experienced by the patient. $\vec{s}^{(t+1)}$ is the state that results from applying parameter $\vec{a}^{(t)}$ in state $\vec{s}^{(t)}$ at time t as governed by the latent dynamical model, f (Figure B.3).

Our Long-Short Term Memory Network (LSTM) neural network, with parameters θ, learns an encoding of sample history, $z^{(t)} = \mathcal{E}_\theta(\mathcal{S}^{(t)})$. This sample history through time, t, is defined as $\mathcal{S}^{(t)} = \langle \vec{s}^{(0)}, \vec{a}^{(0)}, \vec{s}^{(1)}, \ldots, \vec{a}^{(t-1)}, \vec{s}^{(t)} \rangle$ which we refer to as the *meta-state*. Our acquisition function, $Q_\phi : \mathcal{A} \times Z \rightarrow \mathbb{R}$, learns to map a candidate parameter, $\vec{a}$, to a measure of expected informativeness conditioned on the embedding of sample history, $\vec{z}$. This problem setup corresponds to a Partially Observable Markov Decision Process (POMDP), where the observations are our samples, and the state describes the latent dynamics (i.e., the transition function, f) with actions, $\vec{a}$, discount factor, γ, and reward function, R. We do not have access to the observation function, Ω. Similar to [152], we convert this POMDP to an MDP in which we use function approximation to (1) learn a compact representation, $z^{(t)}$, of the history of observations via $\mathcal{E}_\theta$ and leverage this representation to (2) train a history-dependent Q-function, Q_ϕ.

We utilize expected informativeness (i.e., improvement in model accuracy due to the addition of new observations to the training set) as our reward signal for training the network. To determine the decrease in model error, as shown in Equation 6.1, we create a dataset, D^{Test}, by sampling from the known dynamics model, which we have access to during training. The reward signal, $R^{(t)}$, which is defined in Equation 6.2, is the decrease in model error when applying parameter, $\vec{a}^{(t)}$, in state, $\vec{s}^{(t)}$, and experiencing state, $\vec{s}^{(t+1)}$ (i.e., $D_L \cup \langle \vec{s}^{(t)}, \vec{a}^{(t)}, \vec{s}^{(t+1)} \rangle$). Intuitively, a large reward means that we have selected a parameter that greatly decreases the error of the dynamics model, $\hat{f}_\psi$. ψ is the parametrization of $\hat{f}_{\psi^{(t)}}$ at time, t.

$$L_{\psi}(D_L) = \frac{1}{|D_L|} \sum_{i=1}^{|D|} \left(\hat{f}_{\psi} \left(\vec{s}^{(i)}, \vec{a}^{(i)} \right) - \vec{s}^{(i+1)} \right)^2 \tag{6.1}$$

$$R^{(t)} = \frac{\left(L_{\psi^t}(D^{Test}) - L_{\psi^{t-1}}(D^{Test}) \right)}{L_{\psi^t}(D^{Test})} \tag{6.2}$$

In our formulation, chance-constraints allow us to model uncertainty and ensure the probability of failure remains under a certain threshold. Thus, by utilizing a chance-constrained MILP, we can efficiently arrive at a solution for non-convex optimization problems while also providing probabilistic guarantees [153]. We transform each piece-wise term in our acquisition function into a set of integer, linear constraints via the "big M" method [154]. We solve our chance-constrained MILP via linearization techniques discussed in [153, 33]. While limited prior work [155] has explored safety and chance constraints for learning and control, we go beyond this prior work by taking into account the effect that querying a label has on the underlying system's ability to remain in a safe configuration. In our DBS domain, choosing a sequence of unsafe parameters can lead to an ictal state in the brain. As depicted in Figure 6.2, we assume a set of known safe states and we allow the system to deviate from a safe region temporally to gain information provided that the system has a sufficient probability of returning to a safe state. We elaborate on how these safe states are identified and the external validity in Section 6.4. To approximate the uncertainty of the states, we assume our model error comes from a Gaussian distribution with a known mean and variance calculated via the bootstrapping method described in [32].

Finally, we seek to enable our system to generalize beyond a single active learning task (e.g., learning the optimal parameters for a single patient) to a broader class of tasks (i.e., learning the optimal parameters for any patient). We aim to learn this acquisition function without hand-engineering features or heuristics. Therefore, we incorporate meta-learning to train our acquisition function, Q_{ϕ}, and embedding of previously experienced states and actions, $\mathcal{E}_{\theta}$. We train Q_{ϕ} over a *distribution* of optimization problems (i.e., a set of patients

with differing anatomy and disease manifestation) to enable Q_ϕ to generalize to an novel patient.

6.3 Safe Meta-Learning Architecture

Our architecture consists of three key components: (1) an LSTM-based representation of sample history, (2) a meta-learned acquisition function that accurately quantifies expected informativeness, and (3) safety constraints imposed via the linear program. An overview of our architecture is shown in Figure B.3, and is described below.

Policy - Our policy (Equation 6.3) is determined by maximizing both the probability of the system remaining in a safe configuration and expected informativeness along the finite trajectory horizon, $[t, t+T)$. Therefore, our policy selects the set of actions, $\vec{a}^{(t:t+T)}$, which maximizes both safety and expected informativeness. We linearize our objective function following the linearization procedures introduced in [33].

$$\begin{aligned} \vec{a}^{(t:t+T)} &= \pi\Big(\mathcal{E}_\theta(\mathcal{S}^{(t)})\Big) \\ &=_{\vec{a}^{(t:t+T)},\epsilon} \lambda\Big[Q_\phi\left(\vec{a}^{(t:t+T)}, \vec{z}^{(t)}\right)\Big] + (1-\lambda)\Big[1-\epsilon\Big] \end{aligned} \tag{6.3}$$

subject to

$$1 - \epsilon \leq \Pr\Big\{\vec{s}^{(t+T)} - \vec{s}_r(t)_1 \leq \vec{r}\Big\} \tag{6.4}$$

$Q_\phi(\vec{a}^{(t:t+T)}, \vec{z}^{(t)})$ describes the expected informativeness along the trajectory when the set of parameters, $\vec{a}^{(t:t+T)}$, is applied in the context of the sample history encoding, $\vec{z^{(t)}}$. π is the chance-constrained policy which selects the set of parameters that should be applied to the patient to maximize both expected informativeness and safety. The LSTM neural network, $\mathcal{E}_\theta$, maps the sample history, $\mathcal{S}^{(t)} = \langle \vec{s}^{(0:t)}, \vec{a}^{(0:t)} \rangle$, (i.e., previously experienced states

and parameters), to the encoding, $\vec{z}^{(t)}$. λ is a hyper-parameter that allows us to adjust the trade-off between safety and expected informativeness while still guaranteeing a minimum level of safety. Properly balancing λ, as in any multi-criteria optimization problem, requires domain expertise. The estimated probability of remaining in a safe configuration is $1-\epsilon$, where $\epsilon \in [0, \epsilon_{max}]$ and $1-\epsilon_{max}$ is the minimum acceptable safety level. We provide more details on safety in the following section.

Definition of Safety - Next, we detail our safety constraints which are enforced via our MILP. We define a volume of safety, as depicted in Figure 6.2, around a desired safe state, $\vec{s}_r(t)$, and enforce the constraint that the system must be able to return with probability $1-\epsilon$ to a state, $\vec{s}_{t+T}$, at time $t+T$, such that $\vec{s}_{t+T}$ is within this volume of safety. Intuitively, this means that the system will be able to move outside of the volume of safety to gain information but must return to a safe state at time $t+T$. Mathematically, we define this safety constraint in Equation 6.4. $\vec{s}_r$ defines a safe state (e.g., non-cital state for a DBS patient), and $\vec{r}$ is the radius encompassing all known safe system states. The radius of the volume is a hyperparameter defined by the user and requires domain expertise to determine. The closer the user wants the system to remain to the nominal safe state, the smaller the radius should be. $1-\epsilon$ is the probability of remaining in the safe region. This volume can be converted into linear constraints, thus creating a convex optimization problem [33]. Leveraging such a pre-defined safety envelope is consistent with prior work in safe robotics and chance-constrained optimization [156, 157, 158, 159, 160]. By formulating our safety constraints in this way, we can guarantee a minimum probability of safety while simultaneously optimizing for additional safety and expected informativeness. In other words, the MILP will select an parameter that meets the minimum safety requirements, and if possible, will select an even safer parameter than the minimum specified level of safety, all other things being equal.

Meta-learning - To infer the acquisition function, we meta-learn over a distribution of related tasks, which, in our motivating example, consist of a set of diverse DBS patients as shown in Figure B.3. By meta-learning over this distribution, we can construct an acquisition

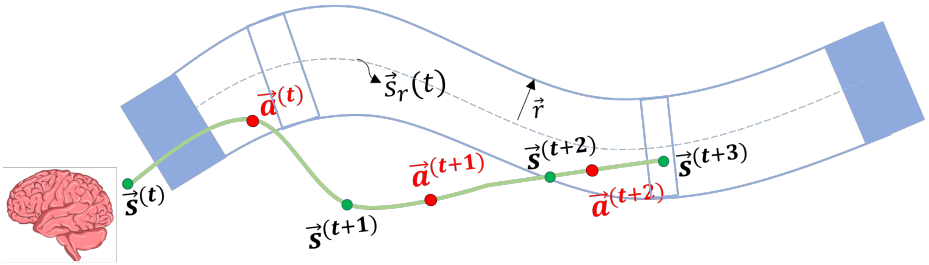


Figure 6.2: This figure depicts the volume of safety, i.e. convex constraints around reference trajectory, $\vec{s}_r(t)$. Action, $\vec{a}^{(t)}$, is an exploratory action, which may bring the system outside of the safe region. Given $\hat{f}_{\psi^{(t)}}$, Safe MetAL ensures the probability that $\vec{a}^{(t+2)}$ returns the system to a safe state is at least $1-\epsilon$.

function that accurately defines the expected informativeness of a parameter when learning the unknown brain model.

The acquisition function, Q_ϕ, is trained via Deep Q-Learning [161] with target network, $Q_{\phi'}$, which has been shown in prior work to improve training stability [162]. The learned acquisition function, Q_ϕ, is utilized by our MILP policy, which selects the optimal actions, $\vec{a}^{(t:T)}$, subject to safety-constraints. The reward, $R^{(t)}$, for applying a set of parameters in a given state is defined as the decrease in the MSE error of the model, $\hat{f}_{\psi^{(t)}}$, achieved by adding training data, $\langle \vec{s}^{(t)}, \vec{a}^{(t)}, \vec{s}^{(t+1)} \rangle$, to D_L, as described in Equation 6.1-Equation 6.2. The Q-function is trained on a set of optimization problems drawn from a distribution of similar black-box functions to minimize the Bellman Residual (Equation 6.5).

$$\mathcal{L}_{\theta,\phi} = \left(R^{(t)} + \gamma Q_{\phi'} \left(\pi \left(\mathcal{E}_\theta(\mathcal{S}^{(t+1)}) \right), \bar{z}^{(t+1)} \right) - Q_\phi(\vec{a}^{(t)}, \bar{z}^{(t)}) \right)^2 \tag{6.5}$$

This Bellman loss of the Q-function is backpropagated through the Q-function in the MILP and through the LSTM encoder, $\mathcal{E}_\theta$. The dynamics model, $\hat{f}_{\psi^{(t)}}$, is retrained with each new set of state-parameter pairs.

Algorithm - Algorithm Figure 2 describes our training procedure. For each episode, we sample from the patient population and limit each episode to the number of time steps, M, tuned to collect enough data to accurately learn the brain model. At each iteration, we select $\vec{a}^{(t)}$ (line 6) via our MILP objective described in Equation 6.3 and apply the parameter setting to observe the resultant state, $\vec{s}^{(t+1)}$ (line 7-8). Our brain model, $\hat{f}_{\psi^{(t)}}$, is retrained by

Algorithm 2 Meta-learning for training

1: Randomly initialize Q_ϕ and $Q_{\phi'}$ with weights $\phi = \phi'$
2: Initialize replay buffer, D
3: **for** episode=1 to N **do**
4: Initialize $\hat{f}_{\psi^{(0)}}$ based on meta-learning distribution
5: **for** t=1 to M **do**
6: Select $\vec{a}^{(t)}$ from Equation 6.3
7: Execute $\vec{a}^{(t)} + \mathcal{N}$ with exploration noise, $\mathcal{N}$
8: Observe state, $\vec{s}^{(t+1)}$
9: $D_L \leftarrow D_L \cup \langle \vec{s}^{(t)}, \vec{a}^{(t)}, \vec{s}^{(t+1)} \rangle$
10: $\psi^{(t)} \leftarrow argmin_\psi L_\psi(D_L)$; observe $R^{(t)}$
11: Update Q_ϕ and $\mathcal{E}_\theta$ via Equation 6.5
12: $Q_{\phi'} \leftarrow \tau Q_\phi + (1-\tau) Q_{\phi'}$
13: **end for**
14: **end for**

minimizing the MSE, as shown in Equation 6.1. After observing the reward (Equation 6.2), we update our Q-function (line 11-12) via a sampled batch of transitions.

Algorithm Figure 3 describes how we perform our online, safe, active learning. Intuitively, our algorithm initializes a new brain model (line 2) to represent the unknown model, and we iteratively sample information rich, safe actions via our MILP policy (line 5), update $\hat{f}_{\psi^{(t)}}$, (line 9) and repeat. We assume at test time that the unknown model comes from the same distribution as the training models.

6.4 Experimental Evaluation - DBS Domain

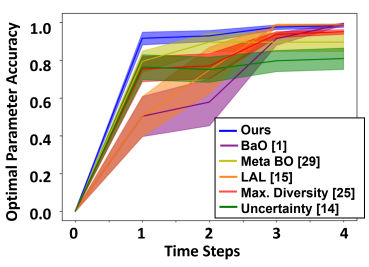
We compare Safe MetAL against several baseline approaches in our DBS domain described below.

DBS is a cutting-edge approach for treating seizure conditions that cannot be controlled via pharmacological methods. Currently, surgeons employ trial-and-error to find control settings that reduce seizures. However, there is no clear mapping from parameter values to

Algorithm 3 Meta-learning for testing

Draw test example from distribution
Initialize $\hat{f}_{\psi^{(0)}}$ based on meta-learning distribution
$D_L \leftarrow \emptyset$
for t=1 to M **do**
  Select $\vec{a}^{(t)}$ according to Equation 6.3
  Execute $\vec{a}^{(t)}$
  Observe state $\vec{s}^{(t+1)}$
  $D_L \leftarrow D_L \cup \langle \vec{s}^{(t)}, \vec{a}^{(t)}, \vec{s}^{(t+1)} \rangle$
  $\psi^{(t)} \leftarrow argmin_{\psi} L_{\psi}(D_L)$
end for

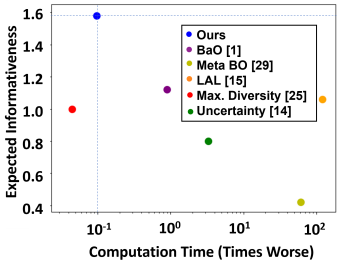
reduction in seizures that applies to all patients, as the optimal parameter settings can depend on placement of the device, the individual anatomy, and other confounding factors. Further, a latent subset of parameters can cause negative side-effects. In keeping with [18], we create simulation environments based on data from six rats where, at each DBS parameter setting, the cognitive function of a rat is measured by a "memory score." Data from each rat is then dissimulated into many digital twins of the rat, creating a population pool over which we can meta-learn. To create these digital twins, we employ a validated *in silico* procedure in which we bootstrap Gaussian Process models trained on *in vivo* data of DBS in rats to create a virtual experimental domain. The task is to determine the DBS parameters (i.e., signal amplitude) in the simulation environments that maximize each rat's memory score (i.e., rat's ability to recall the location of objects) without causing unwanted side effects (e.g., memory deficits or seizures) which occur when the memory score drops below zero. The reward signal utilized by our meta-learner is the percent decrease in error between the predicted and actual optimal parameters. This domain and the established *in silico* evaluation procedure are described further in [18].



(a)

Method	Expected Info	Safety
Our Approach	**1.58** (.022)	**97%** (3.2 %)
BaO [18]	1.00 (.105)	86% (3.0%)
Meta BO [95]	0.43 (.183)	83% (11%)
LAL [94]	1.07 (.098)	88% (2.3%)
Max. Diversity [33]	1.12 (.153)	90% (1.9%)
Uncertainty [32]	0.80 (.241)	**97**% (1.1%)

(b)



(c)

Figure 6.3: This figure depicts our empirical validation in the DBS domain, benchmarking algorithm accuracy per time step (Figure 6.3(a)), overall (Figure 6.3(b)), and vs. computation time (Figure 6.3(c)). The optimal parameter accuracy is defined as $1 - \frac{\vec{a}^* - \hat{\vec{a}}}{\vec{a}^*}$ where $\vec{a}^*$ is the optimal stimulation parameter and $\hat{\vec{a}}$ is the predicted parameter. In Figure 6.3(b) we also report the ground truth safety of our algorithm compared to baselines. The results shown in Figure 6.3(a) comply with the safety results reported in Figure 6.3(b)

6.5 Results

Baseline Comparisons - To demonstrate that meta-learning is a vital component of our framework and produces results superior to prior work, we benchmark against active learning functions, Epistemic Uncertainty [32] and Maximizing Diversity [33]. These active learning functions are linearized and embedded in our safety constrained framework therefore providing a head-to-head comparison between our meta-learned acquisition function and these active learning heuristics. We additionally benchmark against several Bayesian and meta-learning approaches. We empirically validate that Safe MetAL outperforms baselines in the DBS domain in terms of its ability to safely and actively learn latent parameters.

- **Epistemic Uncertainty [32]** - Selects the parameter which maximizes the uncertainy of the model, while also imposing safety constraints via a chance-constrained linear program.
- **Maximizing Diversity [33]** - Selects actions which maximize the difference between previous states and parameters, subject to safety constraints via a chance-constrained linear program.
- **Bayesian Optimization (BaO) [18]** - Developed in previous work for the DBS domain (Figure 6.1) and is based upon a Gaussian Process model which attempts to efficiently determine the optimal parameters.

- **Meta-Bayesian Optimization (Meta BO) [95]** - Meta-learns a Gaussian process prior offline.
- **LAL [94]** - Meta-learns an acquisition function leveraging hand-engineered features.

Active Learning – Results from the DBS domain empirically validate that our algorithm more efficiently learns the optimal parameters (Figure 6.3) compared to baseline approaches.

Safe MetAL selects an action that results in 58% higher expected informativeness and a 267% higher expected informativeness on average compared to our two Bayesian baselines, BaO and Meta BO respectively. Compared to our active learning baselines, Maximizing Diversity and Uncertainty, Safe MetAL performs 41% and 98% better in terms of average expected informativeness respectively. This large increase in expected informativeness that Safe MetAL is able to achieve compared to hand-engineered heuristics, suggests that the meta-learning aspect of Safe MetAL is vital for synthesizing a precise, task-specific acquisition function. Lastly, we show that Safe MetAL outperforms by 47% our meta-learning baseline, LAL, which meta-learns over hand-engineered features. These results demonstrate that our meta-learned embedding is more capable of extracting salient information than the hand-engineered features in LAL.

Safety - Because Safe MetAL is able to more quickly learn the optimal parameter settings, it is also able to ensure safe operation to a greater degree. To empirically validate the safety of each algorithm, we perform a Monte Carlo simulation and determine the percentage of the time that the patient remains in a safe state.

In our DBS domain, Safe MetAL achieves a 6.3% higher guarantee of safety compared to Maximizing Diversity [33] in Figure 6.3. Safe MetAL achieves a 98% greater expected informativeness compared to Uncertainty [32] and achieves an equivalent safety guarantee. In Figure 6.4(b), we show the trade-off between the probability of safety as determined by the MILP and the expected informativeness of an action as a result of adjusting λ. This flexibility allows for greater emphasis on safety in more safety critical domains, whereas in less safety critical domains, these constraints can be relaxed in favor of higher expected

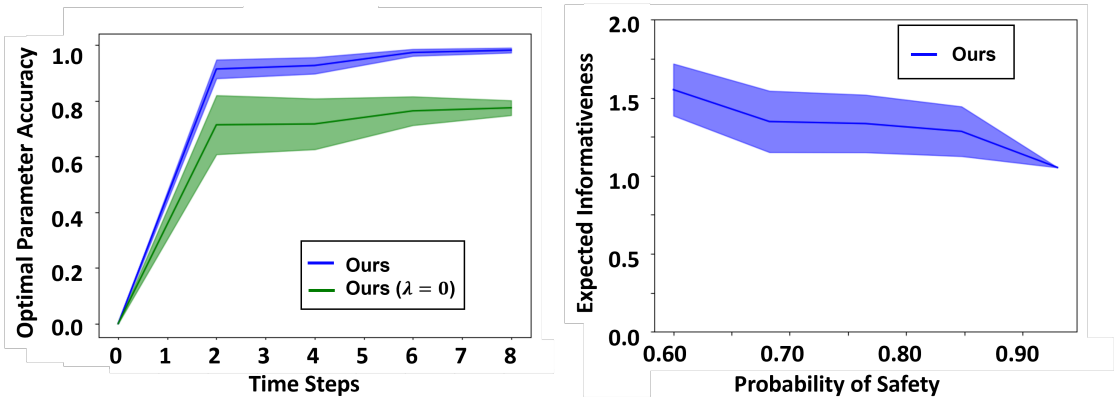
informativeness.

Computation Time - The computation time of active learning algorithms can be of critical importance especially in the time senstive operating room. In our DBS environment (Figure 6.3(c)), BaO has a slight advantage in computation time, but Safe MetAL trades the time for 58% greater expected informativeness. Additionally, Safe MetAL is 68x faster than LAL and 61x faster than Meta BO, our two meta-learning benchmarks.

Ablation Study - To further verify that the meta-learning aspect of Safe MetAL is necessary for achieving high expected informativeness, we perform an ablation study as shown in Figure 6.4. Figure 6.4(a) shows our approach compared to when there is no active learning (i.e., $\lambda = 0$). We show that active learning results in faster selection of the optimal parameter setting. Without active learning, there is little exploration and the optimal parameter setting is never achieved. Additionally, in Figure 6.4(b), we show the tradeoff between safety and expected informativeness when varying λ. We show that expected informativeness decreases very little even with high safety guarantees. By tuning λ we can vary the trade-off between safety and expected informativeness

6.6 Discussion

We present a novel architecture, SafeMetAL, which, unlike previous hand-engineered approaches, leverages sample history to meta-learn a domain-specific acquisition function for safe and efficient control of an unknown system. Through our empirical investigation, we demonstrate that our meta-learned acquisition function operating within a chance-constrained optimization framework outperforms prior work in active learning, meta-learning, and Bayesian optimization [32, 33, 18, 94, 151]. Our approach simultaneously increases expected informativeness while decreasing computation time. Safe MetAL achieves a 41% increase in expected informativeness while decreasing computation time by 20% versus active learning and Bayesian baselines in the DBS domain.



(a) (b)

Figure 6.4: This figure shows the results of our ablation analysis and the trade-off between expected informativeness and safety. In Figure 6.4(a) (DBS domain), we set $\lambda = 0$, meaning there is no active learning and only safety is maximized. This results indicates that our meta-learned acquisition function is an important component to achieve efficient learning. Figure 6.4(b), shows an ablation study, demonstrating the trade-off between expected informativeness and safety when we vary λ. We show that we can tune λ to achieve the desired tradeoff between expected informativeness and safe operation.

To the best of our knowledge, Safe MetAL is the first architecture to meta-learn an acquisition function for active learning embedded within a chance-constrained program for probabilistically safe control. Further, our approach sets a new state of the art over prior work ([18, 32, 94, 33, 151]) for active learning. Our novel, deep learning architecture, offers a unique ability to learn an LSTM-based embedding of sample history while utilizing the power of deep Q-learning to learn a task-specific acquisition function. Safe MetAL is able to optimize both for safety and expected informativeness by embedding our learned acquisition function in a chance constrained optimization framework. With this novel formulation, we demonstrate that Safe MetAL maintains a high probability of safety while also maximizing the expected informativeness based on a learned representation of sample history.

6.7 Limitations and Future Work

Safe MetAL assumes that the safety region is defined by an unchanging volume of safety and that uncertainty over our states is Gaussian. Additonally, Safe MetAL requires data

to meta-learn an acquisition function. However, our results demonstrate that Safe MetAL enables greater expected informativeness and safety when sufficient training data is available. An additional limitation is that the distribution of scenarios from which we meta-learn over requires domain knowledge to determine. Finally, we hypothesize that Safe MetAL's performance depends on the representativeness of the training data, which we will explore further in future work.

6.8 Conclusion

In this paper, we demonstrate Safe MetAL a state-of-the art meta-learning approach for personalized learning with safety constraints. In our approach we 1) accurately quantify domain specific expected informativeness, 2) learn from sample history to improve generalizability and 3) include safety constraints to probabilistically ensure safe sample selection. We demonstrate that our approach achieves a 41% increase in expected informativeness, a 20% speedup in computation time and ensures a high degree of safety across both domains. In this chapter, we advance the field of personalized learning by introducing an approach that is capable of not only learning a personal model of the end-user but is also able to do so in a safe manner.

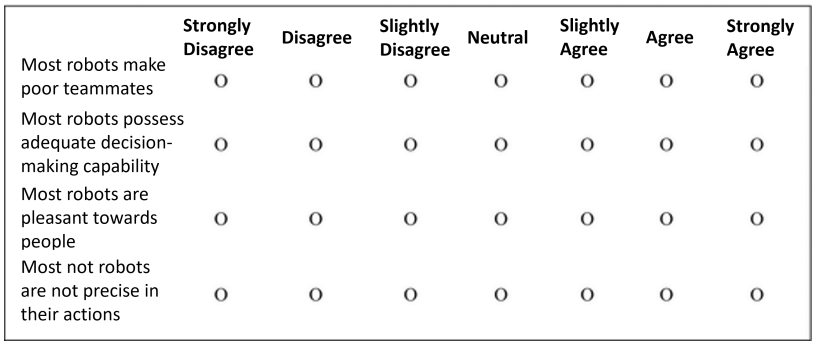
CHAPTER 7

A NOTE ON HUMAN-SUBJECT STUDIES AND LIKERT SCALES

One of the objectives of my thesis is to not only develop data-driven personalized algorithm for human-machine interaction, but to also thoroughly validate these approaches with real human subjects in large human subject studies. To properly validate these approaches, correct measurement tools must be employed and the collected data properly analyzed to ensure statistically sound results.

Likert scales are the most common metric employed in studies to measure a participant's subjective attitude towards a construct. In my work discussed in this thesis, I measure subjective attitudes and characteristics via many different Likert scales including trust in robots [115], Big-Five personality [134], driving style [69], and trust in automation [137]. I rely on these metrics in my work to draw conclusions about the efficacy of our approaches and to determine a human's attitude towards them. However, prior work has shown that Likert scales are often improperly constructed and the data incorrectly analyzed in the field of HRI [163, 113]. For example, researchers often do not employ enough items when constructing a scale and frequently conduct analysis on single Likert items rather than the full scale which can produce inaccurate results [163]. Such improper use and analysis of Likert scales increases the risk for type I errors. To reduce this risk, we as HRI researchers must ensure that we are properly employing and analyzing Likert scales and associated data. Otherwise, we cannot ensure that the algorithmic approaches we develop produce the intended results.

Researchers in the field of psychometrics have conducted extensive analysis on the best way to employ Likert data so as to accurately measure the intended construct [163]. To provide guidance to researchers and decrease the misuse of Likert scales and associated data in my own work, I conducted a review of the psychometric literature on Likert scales



	Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
Most robots make poor teammates	o	o	o	o	o	o	o
Most robots possess adequate decision-making capability	o	o	o	o	o	o	o
Most robots are pleasant towards people	o	o	o	o	o	o	o
Most not robots are not precise in their actions	o	o	o	o	o	o	o

Figure 7.1: This figure illustrates a portion of a balanced Likert scale measuring trust (Courtesy of [166]).

and created recommendations for best practices for researchers [163]. My goal is for this review and the following recommendation to improve my own research practices and ensure that the results presented in this thesis are statistically sound. Additionally, my goal is to aid others in the community in employing best practices with regards to Likert scales. Below are a summary of the recommendations.

7.1 What is a Likert Scale?

Likert scales were created in 1932 by Rensis Likert and were originally designed to scientifically measure attitude [164]. A Likert scale is defined as ”a set of statements (items) offered for a real or hypothetical situation under study” in which an individual must choose their level of agreement [165]. The original response scale for a Likert item ranged from one to five (strongly disagree to strongly agree). A seven-point scale is also common practice. An example Likert scale is shown in Figure Figure 7.1.

Response Format - Confusion often arises around the term ”scale.” A Likert scale does not refer to a single prompt which can be rated on a scale from one to n or ”strongly disagree” to ”strongly agree”. Rather, a Likert scale refers to a set of related prompts or ”items” whose individual scores can be summed to achieve a composite score quantifying a participant’s attitude toward a latent, specific topic [167]. ”Response format” is the more appropriate term when describing the options ranging from ”strongly disagree” to ”strongly agree” [168].

This distinction is important for the following reasons. First, a high degree of measurement error arises when a participant is asked to respond only to a single prompt; however, when asked to respond to multiple prompts, this random measurement error tends to average out. We note that multiple items will reduce random error, but not necessarily systematic error. Second, a single item often addresses only one aspect or dimension of a particular attitude, whereas multiple items can report a more complete picture [169, 170]. Therefore, it is important to distinguish whether there are multiple items in the scale or simply multiple options in the response format. [168] emphasizes the importance of this distinction by stating that the meaning of the term scale "is so central to accurately understanding a Likert scale (and other scales and psychometric principles as well) that it serves as the bedrock and the conceptual, theoretical and empirical baseline from which to address and discuss a number of key misunderstandings, urban legends and research myths."

It is not uncommon in HRI, as well as psychometric literature, for a researcher to incorrectly refer to a response format as a Likert scale. To ground this distinction in an example, Figure Figure 7.1 depicts a Likert scale with four Likert items and a seven-option response format. To avoid such confusion, it is important to be precise when describing a Likert scale, as a five-option response format has a very different meaning from a five-item Likert scale

Distinguishing Between Other Metrics - A psychometric tool should only be labeled as a Likert scale if it complies with the description in this section. Various scales exist that are similar to Likert scales but differ in important ways. For example, a "semantic continuum" consists of a set of semantic differential scales similar to how a Likert scale consists of several Likert items [171]. A semantic continuum differs from a Likert scale in that it utilizes a bipolar scale of antonyms and measures how much of a quality a specific object has. For example, a Likert item may consist of the statement "The robot makes me sad," and the user is prompted to select how much they agree or disagree with the statement. On the other hand, a semantic differential scale will prompt the user to select how the robot makes

them feel, ranging from sad to happy. Multiple semantic differential scales measuring the same attribute can be summed together to form a ”semantic continuum.” While a semantic continuum is appropriate to utilize in many contexts, it has important inherent differences from a Likert scale (for further reading on the differences in data arising from semantic continuums versus Likert scales, please see [172]). For example, semantic continuums are specifically useful for measuring the ”intensity and direction of the meaning of concepts” and have their own set of requirements for design as detailed in [172]. As such, we should be careful to not mislabel one as the other. Additionally, scales such as NASA TLX and SWAT that utilize different or additional methods for calculating composite scores should be distinguished from standard Likert scales via terms such as ”Likert variant” or ”Likert-like” [173, 174].

> *Recommendation - We recommend that HRI researchers be deliberate when describing Likert response formats and scales to avoid confusion and misinterpretation and to only refer to scales that meet the criteria discussed in this section as Likert Scales.*

7.2 Design and Development

Because HRI is a relatively new field, HRI researchers often explore novel problems for which they appropriately need to craft problem-specific scales. However, care must be taken to correctly design and assess the validity of these scales before utilizing them for research. The design of the scale is one of the least agreed upon topics pertaining to Likert questionnaires in the psychometric literature. Disagreement arises around the optimal number of response choices in an item, the ideal number of items that should comprise a scale, whether a scale should be balanced, and whether or not to include a neutral midpoint. The development of the scale also requires rigorous validity and reliability analysis. Below, we address each topic.

Number of Response Options - Rensis Likert himself suggested a five point response format in his seminal work, *A Technique for the Measurement of Attitudes* [164]. However, Likert did not base this decision in theory and rather suggested that variations on this five-point format may be appropriate [164]. Further investigation has yet to provide a consensus on the optimal number of response options comprising a Likert item [175]. [176] found that scales with four or fewer points performed the worst in terms of reliability and that seven to nine points were the most reliable. This finding is backed up by [177] in their investigation of categorization error. [178] demonstrated via simulation that the more points a response contains, the more closely it approximates interval data and therefore recommended an 11-point response format.

This line of reasoning may lead one to believe that one should dramatically increase the number of response points to more accurately measure a construct. However, just because the data may more closely approximate interval data does not mean increasing the number of response points monotonically increases the ability to measure a subject's attitude. A larger number of response options may require a higher mental effort by the participant, thus reducing the quality of the response [179, 180]. For example, [179] conducted a study that suggested that response quality decreased above eleven response options. [181] also investigated the optimal number of response options and found that no further psychometric advantages were obtained once the number of response options rose above six and [180] suggested based on study results that the optimal number is between four and six.

> *Recommendation - As a general rule-of-thumb, we recommend the number of response options be between five and nine due to the declining gains with more than ten and lack of precision with less than five. However, if the study involves a large cognitive load or lengthy surveys, the researcher may want to err on the side of fewer response items to mitigate participant fatigue [176].*

Response Format Label - By the formal definition, a Likert scale response format should be labeled from "strongly disagree" to "strongly agree" [164]. Although there is little evidence

in the psychometric literature to suggest that this choice of label is superior to other choices, other response format labels have not been widely studied and therefore are not as well understood. Furthermore, a review conducted by [182] suggests that the response format label may have an impact on data quality and interpretation.

There is further debate about the label of the midpoint (see below for a discussion about inclusion versus omission of a midpoint). Likert's original scale utilized the label "undecided" for the midpoint [164]. However, researchers have suggested that either "neutral" or "neither agree nor disagree" are better alternative to "undecided" as "undecided" may represent an absence of opinion and therefore not comply with the ordinal nature of the response format [183].

Prior work [184] has also investigated the labeling of Smiley-o-Meter scales which are Likert-like scales commonly employed in research with children. The standard Smiley-o-Meter utilizes smiley faces as labels, typically ranging from sad to happy. Hall and Hume conducted several studies with various response labels and found that children rarely selected the negative ratings, perhaps because children are tuned to more positive attitudes [184]. To solve this issue, the researchers created the Five Degree of Happiness scale which utilizes varying degrees of happy faces for the response labels which produced higher quality responses by encouraging the use of all scale points in studies with children.

> *Recommendation - We recommend that authors adhere to the "strongly disagree" to "strongly agree" response format label when possible, as this has been thoroughly validated. Further, we recommend that authors utilize either "neutral" or "neither agree nor disagree" when labeling a midpoint to maintain the ordinal nature of the scale. When deviating from this label, we recommend that authors instead refer to their scale as "Likert-like" to differentiate it from the classical Likert scale. When soliciting responses from children, utilize the Five Degrees of Happiness scale [184]*

Neutral Midpoint - Another point of contention which influences the response format of

a scale is whether or not to include a neutral midpoint. Likert, with his five-point scale, included a neutral, "undecided" option for participants who did not wish to take a positive or negative stance [164]. Some argue that the inclusion of a neutral midpoint provides more accurate data because it is entirely possible that a participant may not have a positive or negative opinion about the construct in question. Studies have shown that including a neutral option can improve reliability in other, similar scales [185, 186, 165, 187]. Furthermore, the lack of a neutral option precludes the participant from voicing an indifferent opinion, thus forcing them to pick a side which they does not agree with.

On the other hand, a neutral midpoint may result in users "satisficing" (i.e., choosing the option that may not be the most accurate to avoid extra cognitive strain resulting in an over-representation at the midpoint) [188]. The authors in [189] argue that "...the midpoint should be offered on obscure topics, where many respondents will have no basis for choice, but omitted on controversial topics, where social desirability is uppermost in respondents' minds."

> *Recommendation - We adopt the recommendation of [189], which suggests that HRI researchers utilize their best judgement based on the context of use when deciding the merits of including a neutral option in their response format. For example, if the authors are conducting a pre-trust survey to gauge a baseline level of trust before the participant has interacted with the robot, they may want to include a neutral option since some participants, especially those unfamiliar with robots, may not truly have a good sense of their own trust in robots. A neutral option would allow participants to present this sentiment. However, if a survey is being utilized to assess trust after a participant has interacted with a robot, the researchers may want to remove the neutral option, based on the notion that participants should have developed a sense of either trust or distrust after the interaction. Nonetheless, there may be cases when "neutral" truly is appropriate, which is why we argue in favor of researcher discretion [189].*

Overall Response Format Design - The number of response options and the response format labels are intrinsically linked. The number of response options inevitably influences the choice of response labels. The more response options, the more difficult it is to assign a label to each option. Typically scales with many response options must rely on anchor labels with either number labels or no labels for intermediate options. Prior work has investigated the differences that arise in fully labeled versus partially labeled scales as well as the effect of gradation of (dis)agreement (e.g., a five-point scale has two gradations whereas a seven-point has three) when labeling the response scale [190]. Weijters et al. found that a fully labeled scale led to higher quality responses [190]. Thus, the authors recommend in situations of opinion measurement and scale development to utilize either a five-point or seven-point *fully labeled* response format. These findings are supported in other studies which demonstrate that a fully labeled scale produces higher reliability [191, 192].

> *Recommendation - In alignment with our above recommendations on number of response options and response format labels and the recommendations provided in [190], we recommend that authors utilize a five-point or seven-point fully labeled response format to achieve high-quality responses. In a five-point response format, authors should label the options "strongly disagree," "disagree," either "neutral" or "neither agree nor disagree," "agree," and "strongly agree." In a seven-point response format, authors should label the options "strongly disagree", "disagree", "slightly disagree," either "neutral" or "neither agree nor disagree," "strongly agree," "agree," and "slightly agree." However, we recognize that little research has been conducted on the exact nature of the response format labels and therefore, we provide this recommendation only as a soft guideline.*

Number of Items - The next point of contention we address is the ideal number of Likert items in a scale. In his original formulation, Likert stated that multiple questions were

imperative to capture the various dimensions of a multi-faceted attitude. Based on Likert's formulation, the individual scores are to be summed to achieve a composite score that provides a more reliable and complete representation of a subject's attitude [169, 170].

Yet, in practice it is not uncommon for a single item to be used in HRI research due to the efficiency that such a short scale provides. Research into the appropriateness of single item scales has been extensively studied in marketing and psychometric literature [193]. For example, [193] investigated the use of a single-item scale for measuring a construct concluding that a single-item scale is only sufficient for simple, uni-dimensional, unambiguous objects.

Multi-item scales on the other hand are "suitable for measuring latent characteristics with many facets." [194] proposed a procedure for developing scales for evaluating marketing constructs and suggested that if the object of interest is concrete and singular, such as how much an individual likes a specific product, then a single item is sufficient. However, if the construct is more abstract and complex, such as measuring the trust an individual has for robots, then a multi-item scale is warranted. This line of reasoning is supported by [195, 196, 197]. As to the exact number of items, [196] demonstrated via simulation that at least four items are necessary for evaluation of internal consistency of the scale. However, as suggested by [**Willits**], one should be cautious of including too many items, as a large scale may result in higher refusal rates (i.e., more unanswered questions).

> *Recommendation - Due to the complexity of attributes most often measured in the field of HRI (e.g., trust, sociability, usability, etc.), we recommend that researchers in the HRI community utilize multi-item scales with at least four items. The total number of items again is left to the discretion of the researcher and may depend on the time constraints and the workload that the participant is already facing. Because an average person takes two to three seconds to answer a Likert item and individuals are more likely to make mistakes or "satisfy" after several minutes, we recommend surveys not be longer than 40 items [198].*

Recall that this recommendation for the number of "Likert Items" is distinct from our recommendation regarding the number of "response options," which we recommend generally be between five and nine options, as noted previously.

Scale Balance - The last aspect of scale design which we will discuss is that of balance. The question of whether the items within a scale should be balanced, i.e., there should be a parity of positive and negative statements, is one less often addressed in literature. It is believed that balancing the questionnaire can help to negate acquiescence bias, which is the phenomenon in which participants have a stronger tendency to agree with a statement presented to them by a researcher. Likert [164] advocated that scales should consist of both positive and negative statements. Many textbooks, such as [199], also state that scales should be balanced. Perhaps the most compelling evidence that balance is an important factor when developing Likert scales is provided by [200]. The authors in [200] conducted a study in which they asked participants to respond to a positively worded question to which 60% of participants agreed. They asked the same question but rephrased in a negative way and again, 60% of participants agreed. This study reveals the extent to which acquiescence bias can sway participants to answer in a particular way that is not always representative of their true feelings.

One would find this evidence to be sufficiently compelling to recommend scale balance; however, this debate is not so easily settled. Recent work suggests that although including both positively and negatively worded items reduces the effects of acquiescence bias, it may have a negative impact on the construct validity (i.e., if the scale adequately measures the construct of interest) of the scale [201, 202]. This result may be due to the fact that a negatively worded item is not a true opposite of a positively worded item. Therefore, reversing the scores of the negatively worded items and summing may have an impact on the dimensionality of the scale due to the confusion that reversed items cause [203, 204].

Recommendation - Because of a lack of consensus and the problems arising from

both approaches, we do not provide a concrete recommendation to researchers about scale balance.

Validity and Reliability of Likert Scales - The reliability (i.e., the scale gives repeatable results for the same participant) and the validity (i.e., the scale measures what is intended) of the scale are both contingent on the guidelines listed above. For example, [170] found that a single item scale decreased reliability, and as discussed by [205], using scales with five to seven response options improves reliability and validity. Additionally, Likert's original work states that the prompts of a Likert scale should all be related to a specific attitude (e.g., sociability) and should be designed to measure each aspect of the construct. Each item should be written in clear, concise language and should measure only one idea [206, 164]. This formulation helps to ensure the reliability and the validity of the scale. Therefore, to improve validity and reliability, researchers should closely adhere to the above recommendations when designing Likert scales.

Even if these guidelines are followed, ensuring the reliability and validity of a scale is not a simple task. Rigorous analysis and repeated studies should be conducted to confirm the legitimacy of the scale before use. When designing a scale, an initial pool of items (two times to five times the size of the desired size of the final scale) should be created [207]. Items should be derived from theory and prior work. Content validity of each item should be verified by experts in the field. Items can then be eliminated via factor analysis and measures of internal validity to form the final scale. Common methods for item reduction include the Classical Test Theory (CTT) and Item Response Theory (IRT) which rely on item difficulty index, discrimination index, inter-item and item-total correlations, and distractor efficiency analysis to determine the best items in the pool [208, 209, 207, 210]. If, after CTT or IRT has been applied to the scale, the number of items are less than the recommended minimum of four items, the researchers should create additional items based on theory and expert knowledge.

Recommendation - Due to the complex nature of scale design, we recommend

that researchers utilize well-established and verified scales provided in literature when possible. Many common constructs measured in the field of HRI can be measured with already validated scales such as the "HRI Trust Scale" for human-robot trust [211] or the RoSAS scale for perceived sociability [212]. This practice will reduce the prevalence of employing poorly designed scales.

Internal Consistency and Dimensionality - A poorly formed scale may result in data that does not assess the intended hypothesis. Thus, before a statistical test is applied to a Likert scale, it is best practice to test the quality of the scale. Cronbach's alpha is one method by which to measure the internal consistency of a scale (i.e., how closely related a set of items are). A Cronbach's alpha of 0.7 is typically considered an acceptable level for inter-item reliability [213]. If the items contains few response options or the data is skewed, another method, such as ordinal alpha, should be employed [214]. Cronbach's alpha alone does not ensure the reliability of a scale. For example, a scale consisting of unrelated prompts may achieve a high Cronbach's alpha for other underlying reasons or simply because Cronbach's alpha can increase as the number of items in the scale increases [215, 216]. Therefore, it is also good practice to utilize a test-retest method in which the scale is tested within the same population across multiple points in time in addition to reporting Cronbach's alpha [217]. Furthermore, recent work has suggested that other internal consistency metrics such as McDonald's omega coefficient, ω, may provide better estimates of reliability [218, 219]. For further discussion on this topic, please reference [218].

While Cronbach's alpha and other reliability tests are important metrics, a full item factor analysis (IFA) should be conducted to better understand the dimensionality of a scale. A scale can show internal consistency, but this does not mean it is uni-dimensional. On the other hand, a factor analysis is a statistical method to test whether a set of items measure the same attribute and whether or not the scale is uni-dimensional. Factor analysis thus provides a more robust metric to assess the scale quality [220].

Additionally, factor analysis is crucial in scale development to determine which items

load on each factor. A factor, in this context, describes a latent variable. For example, in the RoSAS scale, a tool commonly used in HRI research, these latent variables are warmth, competence, and discomfort [212]. During scale development, factor eigenvalues, derived from Factor Analysis (FA), are utilized to determine the importance of each factor. Factors with an eigenvalue greater than one are retained. Factor loading values are commonly employed to determine which items load onto each factor. It is recommended to retain items that have a factor loading of above 0.4 because these items explain more than 10% of the variance in the data [221, 207].

> *Recommendation - If researchers choose to create their own scales rather than employing well-established scales from prior work, a thorough analysis of the internal consistency and dimensionality of new scales should be conducted before deployment. Factoring loading values for individual items should be at least 0.4 and factors with eigenvalues greater than one should be retained. For in-depth instructions on how best to construct Likert scales from the ground up, please see [222, 223, 207].*

7.3 Statistical Tests

Once a scale is designed and its validity statistically verified, it is important that correct statistical tests are applied to the response data obtained from the scale. Another fiercely debated topic is whether data derived from single Likert items can be analyzed with parametric tests. We want to be clear that this controversy is not over the data type produced by Likert items but whether parametric tests can be applied to ordinal data.

Ordinal versus Interval - Previous work has demonstrated that a single Likert item is an example of ordinal data and that the response numbers are generally not perceived as being equidistant by respondents [224]. Because the numbers of a scale for Likert items represent ordered categories but are not necessarily spaced at equivalent intervals, there

is not a notion of distance between descriptors on a Likert response format [225]. For example, the difference between ”agree” and ”strongly agree” is not necessarily equivalent to the difference between ”disagree” and ”strongly disagree.” Thus, a Likert item does not produce interval data [226]. While it has been speculated that a large-enough response scale can approximate interval data, Likert response scales rarely contain more than 11 response points [227, 178].

> *Recommendation - Because a Likert item represents ordinal data, parametric descriptive statistics, such as mean and standard deviation, are not the most appropriate metric when applied to individual Likert items. Mode, median, range, and skewness are better to report.*

Parametric versus Non-Parametric - The question now becomes, given the ordinal nature of individual Likert items, is it appropriate to apply parametric tests to such data? A famous study by [228] showed that the F test is very robust to violation of data type assumptions and that single items can be analyzed with a parametric test if there are a sufficient number of response points. [224] demonstrates through simulation that ANOVA is appropriate when the single-item Likert data is symmetric but that Kruskal-Wallis should be used for skewed Likert item data. [177] also found that skew in the data results in unacceptably high errors when the data is assumed to be interval. [229] compared the use of the t-test versus the Wilcoxon signed rank test on Likert items and found that the t-test resulted in a higher Type I error rate for small sample sizes between 5 and 15. [230] made a similar comparison and also found that Wilcoxon rank-sum outperformed the t-test in terms of Type I error rates. As demonstrated by these studies, the field has yet to reach a clear consensus on whether parametric tests are appropriate, and if so when, for single Likert item data.

Likert scale data (i.e., data derived from summing Likert items) can be analyzed via parametric tests with more confidence. [228] showed that the F test can be used to analyze full Likert scale data without any significant, negative impact to Type I or Type II error rates

as long as the assumption of equivalence of variance holds. Furthermore, [231] showed that Likert scale data is both interval and linear. Therefore, parametric tests, such as analysis of variance (ANOVA) or t-test, can be used on full Likert scales as long as the appropriate assumptions hold.

> *Recommendation - Because studies are inconclusive as to whether parametric tests are appropriate for ordinal data, we recommend that researchers err on the conservative side and utilize non-parametric tests when analyzing single Likert items. However, we also recommend that HRI researchers avoid performing statistical analysis on single Likert items altogether. As [168] so eloquently states, "one item a scale doth not make." A single item is unlikely to be the best measure for the complex constructs that are of interest in HRI research as discussed in Section 7.2. Therefore is best to avoid the ordinal vs. interval controversy altogether and instead perform analysis on a multi-item scale since Likert scales can be safely analyzed with parametric tests if appropriate assumptions are met. If a researcher does choose to analyze an individual item, they should clearly state they are doing so and acknowledge possible implications. At the very least, it is recommended to test for skewness.*

Post-hoc Corrections - The importance of performing proper post-hoc corrections and testing for assumptions applies to all data and is not specific to Likert data. Nevertheless, they are important considerations when analyzing Likert data and are often incorrectly applied in HRI papers.

As the number of statistical tests conducted on a set of data increases, the chances of randomly finding statistical significance increases accordingly even if there is no true significance in the data [232]. Therefore, when a statistical test is applied to multiple dependent variables that test for the same hypothesis, a post-hoc correction should be applied. Such a scenario arises frequently when a statistical analysis is applied to individual items

in a Likert scale [168]. In 2006, [233] conducted a study investigating whether individuals born under a certain astrological sign were more likely to be hospitalized for a certain diagnosis. The authors tested for over 200 diseases and found that Leos had a statistically higher probability of being hospitalized for gastrointestinal hemorrhage and Sagittarians had a statistically higher probability of a fractured humerus. This study demonstrated the heightened risk of Type I error that occurs when no post-hoc correction is applied.

There is controversy as to which post-hoc correction is best. [234] suggests applying the Bonferonni correction when only several comparisons are performed, i.e., ten or less. The authors recommend employing a different correction such as Tukey or Scheffé with more than ten comparisons to avoid the increased risk of Type II errors that stems from the conservative nature of the Bonferonni correction. The authors of [235] suggest that researchers should, instead of performing post-hoc correction, focus on reporting effect size and confidence intervals, such as Pearson's r.

> *Recommendation - Because of the danger that comes with performing many statistical tests without predefined comparisons, we recommend that researchers always perform the proper post-hoc corrections. Due to the increased risk of Type II error that some post-hoc tests pose, we encourage researchers to also report the effect size and confidence interval to provide a more informative and holistic view of the results. In general, we recommend against pair-wise comparisons performed on individual Likert items for reasons already discussed.*

Test Assumptions - Most statistical tests require certain assumptions to be met. For example, an ANOVA assumes that the residuals are normally distributed (normality) and the variances of the residuals are equal (homoscedasticity) [236]. Tests to ensure these conditions are met include the Shapiro-Wilk test for normality and Levene's test for homoscedasticity [237]. [228] argues that even when assumptions of parametric tests are violated, in certain situations, the test can still be safely applied. However, [238] counters [228] and contends that [228]

failed to take into account the power of parametric tests under various population shapes and that these results should not be trusted.

> *Recommendation - To navigate this controversy, we suggest that researchers err on the conservative side and always test for the assumptions of the test to reduce the risk of Type I errors. If the data violates the assumptions, and the researchers decide to utilize the test despite this, they should report the assumptions of the test that have not been met and the level to which the assumptions are violated.*

7.4 Conclusion

A majority of published HRI papers rely on Likert data to gain insight into how humans perceive and interact with robots, leading Likert questionnaires to be a fundamental part of HRI studies [113]. I conducted a review of the psychometric literature focused on incorrect design and statistical testing of Likert scales and associated data to explore the implications of these infractions. These guidelines have aided me in following best practices in my own research. It is my hope that our recommendations are taken into consideration and that HRI researchers, authors, and reviewers employ best practices when addressing Likert data.

CHAPTER 8
CONCLUSION

Throughout this thesis, I have made contributions to data-driven personalization techniques for human-machine interaction. This thesis sets out to develop and validate novel personalized frameworks that 1) learn from both the population as well as the individual end-user, 2) do so in a variety of domains, and 3) consider safety of the end-user. Throughout this thesis, I have developed novel approaches which have accomplished these three objectives; in large human-subject studies with over 330 total participants, I demonstrated that these frameworks produce personalized behaviors and improved the machine's ability to interact with the end-user. These frameworks not only objectively perform better but are also viewed as more likeable and trustworthy by participants. The following section provides a summary of each of these methods.

8.1 Mutual Information Driven Meta-Learning from Demonstration

I have presented MIND MELD, a novel framework for learning from suboptimal and heterogeneous demonstrators. My approach learns a personalized embedding that describes the way in which an individual is suboptimal via data collected through calibration tasks. My experimental results show that our approach is capable of improving an agent's ability to learn from novice end-users. I show that our approach both objectively and subjectively outperforms human-centric and robot-centric LfD approaches.

8.2 Personalized Teaching via Reciprocal Mutual Information Driven Meta-Learning from Demonstration

I introduced Reciprocal MIND MELD which expanded upon our MIND MELD approach to provide personalized instructions to suboptimal demonstrators. Reciprocal MIND MELD provides actionable robotic feedback based upon the location of a demonstrator's personalized embedding to improve upon their quality of demonstrations. Across three human-subject experiments, I demonstrated that our approach improves upon a demonstrator's ability to provide high-quality feedback. I also showed that Reciprocal MIND MELD is capable of predicting how a demonstrator's teaching abilities change over time and adapting feedback accordingly.

8.3 Manipulating Autonomous Vehicle Embedding Region for Individuals' Comfort

In my next work, I presented an approach for learning about and accounting for an individual end-user's preferences in an AV domain. I introduced MAVERIC, a personalized framework that learns to mimic an end-user's own driving style and adjust the level of aggression based upon individual characteristics. I experimentally show that our approach is capable of mimicking driving style as well as producing more cautious and aggressive behavior. Additionally, I find that certain latent factors including personality, perceived similarity, and self-reported high-velocity driving style impact the effect of homophily.

8.4 Safe Meta-Active Learning for Deep Brain Stimulation

Finally, I presented Safe MetAL, a personalized algorithm for efficiently selecting the optimal parameter settings and reasoning about the safety of a DBS patient in a healthcare setting. In this work, I combine a deep-learning based acquisition function with safety constraints to safely and efficiently learn the optimal parameter settings for a DBS patient. My experimental results show that our approach outperforms prior work in terms of information

gain, safety, and computation time.

CHAPTER 9
FUTURE WORK

While my work discussed in this thesis demonstrates the potential for data-driven personalization approaches to improve upon human-machine interaction, there is still much untapped potential left to be explored. Below I discuss some of the possible avenues for future work. I propose future directions related to personalization in LfD and discuss future work in learning to transfer knowledge about an individual demonstrator from one domain to another. I additionally propose future steps for learning to differentiate between preference and suboptimality in LfD. I then discuss how Reciprocal MIND MELD can be employed as a general tutoring framework to provide personalized coaching to correct for end-user suboptimality in a variety of domains. I also suggest future steps for improving upon my MAVERIC framework. Lastly, I discuss future work in evaluating Safe MetAL on a target population and how personalized frameworks can be used to improve patient care on a broad scale.

9.1 Transferring Understanding of Suboptimality Across Domains via MIND MELD

In my MIND MELD work, I introduced an approach for learning about the way in which an individual demonstrator is suboptimal within a specific domain. However, because end-users will likely need to teach a variety of different robotic platforms, robots should be capable of transferring the learned knowledge about an individual's suboptimal tendencies from one domain to another. For example, an individual may need to teach a mobile robot a new route in their home but also teach a dish-washing robot how to move dishes from the cabinet to sink. In both domains, the suboptimality of the individual's demonstrations must be taken into account to effectively learn from the human. In future work, I propose an approach called Domain Agnostic MIND MELD which learns to map the personalized embedding

learned in one domain to an embedding describing the demonstrator's suboptimality in another domain.

To learn this mapping, I propose to collect human calibration data to learn an individual's personalized embedding, $w_d^{(p)}$, in one domain (e.g., a driving domain). The same participants will also perform calibration tasks in a different domain (e.g., 7-DOF arm domain) to learn their embedding, $w_a^{(p)}$. By collecting calibration data from a set of participants, we can learn a mapping of the personalized embeddings, $w_d^{(p)} = m_{a \to d}(w_a^{(p)}, \vec{c}^{(p)})$ from one domain to another, conditioned on relevant covariates, $\vec{c}$ (e.g., experience with video games, experience teleoperating a robot, etc.). This transferred embedding will be used to improve upon suboptimal labels in the new domain, thereby enhancing the robot's ability to learn from suboptimal human feedback across disparate domains.

9.2 Differentiating Between Preference and Suboptimality

In my work discussed in this thesis, I investigated the question of heterogeneity with regards to suboptimality in LfD. However, humans are also heterogeneous with respect to their preferences for how new tasks should be performed. Prior work has investigated how to extract different preferences from heterogeneous demonstrators. For example, Chen et al. proposed Fast Life-Long Adaptive Inverse Reinforcement Learning or FLAIR [15]. FLAIR distills common knowledge from demonstrators while still preserving demonstrator preferences. By training FLAIR on a distribution of demonstrators representing the range of preference for end-users, the authors show that we can gain an understanding of the set of prototypical strategies employed by demonstrators.

While this work is able to successfully learn preference based policies for heterogeneous end-users, FLAIR assumes that all demonstrators are optimal when demonstrating their preferences. However, as I illustrated in our MIND MELD work, this assumption often does not hold. Therefore, a robot must be capable of distinguishing between heterogeneity that is a consequence of suoptimality versus heterogeneity due to differing preferences

to effectively learn the intended policy. Yet it can be difficult to elicit which aspects of heterogeneity are due to suboptimality as opposed to preference.

One way to approach this problem is for the robot to actively query the user about their intentions to gain further insight into the goal of the end-user. This information enables the robot to determine what aspects of heterogeneity are related to preference and which are related to suboptimality. The robot can then learn to preserve the aspects of demonstrations related to preference and correct for those related to suboptimality. In future work, I plan to achieve this goal by combining FLAIR and MIND MELD. By doing so, the machine can learn to differentiate between suboptimality and preferences and avoid the incorrect assumptions that 1) there is an optimal way to complete the task, and 2) human demonstrators are optimal.

9.3 Personalized Tutor Via Reciprocal MIND MELD

In this thesis, I introduced Reciprocal MIND MELD as a personalized teaching framework for LfD. However, Reciprocal MIND MELD has the potential to quantify suboptimality and provide robotic feedback in a multitude of domains unrelated to LfD - i.e., to be used as a general coaching and tutoring framework. By learning about the way in which a student is suboptimal, our Reciprocal MIND MELD approach could provide personalized instructions in a variety of different domains.

For example, Reciprocal MIND MELD could be utilized in a domain such as table tennis to teach a human how to improve upon their table tennis stroke. By observing the human playing table tennis and collecting data about the human's stroke, we could employ a similar framework as Reciprocal MIND MELD to learn a personalized embedding which captures information about the way in which the human is suboptimal. In this approach, the robot would then utilize this information to provide feedback to the human.

This aim provides many different avenues for future research. For example, a domain such as table tennis poses different challenges compared to a driving simulator domain. The

dimensions of suboptimality are likely much more complex and difficult to communicate to the end-user. Additionally, players will exhibit differing playing styles which should be accounted for when providing feedback. Because of the complexity of suboptimality, verbal feedback may not be the best option. We can instead take inspiration for how human coaches communicate about suboptimality in complex physical tasks. Prior work has shown that teaching by example is a powerful tool for improving upon suboptimality in physical skills. For a robot to teach by example, it must have a notion of what better and worse looks like and be able to translate this notion into physical movement.

Another challenge in table tennis is that a robotic coach must consider differing playing styles. One individual's stroke may differ from another's but that doesn't necessarily mean that one stroke is superior. Thus, in such a domain, we want to ensure that a robot can provide examples of what a better demonstration looks like while preserving the individual's style of play. To address these challenges, I take inspiration from our MAVERIC work in which we learned the aggressive gradient in embedding space. In a similar manner, in a table tennis domain, we can learn the skill gradient within embedding space. Moving an embedding in the positive direction of the skill gradient would provide an example of what a better stroke looks like while preserving style. Moving an embedding in the direction perpendicular to the skill gradient would maintain skill while varying style. This framework would allow a robot to provide examples of what better and worse strokes look like while maintaining the style of the individual player. By doing so, the robot could coach a table tennis player via examples.

9.4 Establishing Causal Relationships and Quantifying Magnitude of Embedding Shift to Optimize Driving Style

My MAVERIC framework introduced an approach for mimicking an end-user's driving style and increasing and decreasing the aggressiveness relative to one's own driving style. In this work, we found that various latent factors impacted an individual's preference for a

driving style different from one's own. However, to optimize driving style for an individual, we must determine the amount by which to shift an end-user's personalized embedding in the direction of the aggressive gradient. Therefore, one avenue for future work is to study and quantify the relationship between these subjective factors and preference for style. To do this, in future work I aim to conduct a human-subject study to determine the optimal aggressive driving style for each participant and then utilize this information to quantify the relationship between the optimal style and latent subjective factors.

Furthermore, there are likely many different situational and environmental factors that influence an individual's preference for a specific driving style. For example, inclement weather will likely impact the optimal driving style and certain individuals may prefer a less aggressive driving style in these situations. On the other hand, if an individual is late for an appointment, they may prefer their AV to drive in a more aggressive manner. In future work, I am aim to explore how we can adapt driving style on-the-fly to account for these different variables.

9.5 Evaluation of Safe MetAL with Target Population

One of the goals of my research is to develop data-driven approaches for human-machine interaction while also rigorously validating these approaches in large human subject studies. Due to the limitation of deploying novel frameworks in a medical setting with patient populations, we have only demonstrated our Safe MetAL work in a simulation based on rat models. In future work, I aim to validate our approach with human patients with Parkinson's disease. We aim to collect a dataset of DBS stimulation parameters and metrics related to symptom improvement and side-effects of Parkinson's patients. By training our Safe MetAL architecture on this dataset, we hypothesize that we will be able to efficiently determine the optimal DBS parameter setting for an individual patient and demonstrate the efficacy of our approach with the target population.

There are many interesting challenges that arise when deploying a machine learning

framework in a clinical setting. For example, our framework will likely need to learn from incomplete datasets with only partial parameter sweeps for each patient. One technique to solve this problem is to employ Generative Adversarial Networks (GANs) to artificially increase the size of the dataset for each patient. Another challenge is that our safety constraints will have to be carefully designed and tested to ensure patients do not experience negative side-effects. By addressing these challenges and deploying Safe MetAL in a clinical setting, we will move one step closer to improving the lives of patients via our Safe MetAL approach

9.6 Personalized Learning Applied to Healthcare

A patient's response to medication and treatments is often hard to predict, making it difficult for physicians to determine the right course of action for an individual patient. My Safe MetAL framework addressed this problem for DBS patients by predicting the optimal personalized parameter setting. Yet, DBS is only the tip of the iceberg when it comes to data-driven systems supporting personalized patient care. With modern electronic patient databases, we now have the resources to develop data-driven, personalized care systems to aid physicians in diagnosis as well as provide real time feedback to patients about how they can adjust their diet and lifestyle habits to improve their health.

The frameworks developed in this thesis serve as a strong starting point for developing personalized healthcare approaches that can provide unique insight into a patient's health. By learning a personalized embedding from large datasets of patient information, we can utilize this embedding to gain insight into informative biomarkers such as the gut microbiome which can be otherwise difficult to measure. Such approaches can be particularly useful in healthcare paradigms in which individual responses vary greatly due to latent factors that can be difficult for physicians to understand.

For example, the postprandial glucose response from the same meal varies greatly across individuals. This variation, which is likely due to a variety of latent variable (e.g, the makeup

of the gut microbiome, physical activity, etc.), makes it difficult for diabetic patients to optimize their diet and effectively reduce the magnitude of their glucose response [239]. Utilizing the techniques introduced in this thesis, we can learn a personalized embedding that captures information about these latent variables and produces a more accurate prediction of glucose response, thus allowing patients to make better dietary choices. This serves as one example of many in which personalization techniques have the potential to improve the lives of patients.

9.7 Ethics of Personalization

To effectively support humans, machines must be capable of recognizing individual desires, abilities, and characteristics and adapt to account for differences across individuals. However, personalization does not come without a cost. In many domains, for robots to effectively personalize their behavior to match the needs of end-users, the robot must solicit often private and intimate information about an end-user so as to optimize the interaction. However, not all end-users may be comfortable sharing this information, especially if they do not have insight into why the robot is requesting it. As HRI researchers, we have the responsibility of ensuring that the robots we create do not infringe upon the privacy rights of end-users.

In future work, I aim to conduct a study investigating questions related to sharing private information for personalization purposes. Specifically, I propose to investigate the impact of domain, robot embodiment, nature of personal information requested, and the role of explanations with regards to end-user willingness to share information with a robot. My goal is to provide guidance for HRI researchers who are conducting work in personalization by examining the factors that may impact acceptance of personalized robots. Furthermore, to reduce the amount of sensitive information that a machine must collect from an end-user, I aim to investigate methods for active learning and uncertainty quantification with regards to learning a model of the end-user.

9.8 Developing a Unified Personalized Framework

My thesis and my proposed directions for future work are stepping stones to developing a unified framework to enable a machine to understand its end-user and utilize this information to optimize interactions in many domains and across time. An end-user's personality, prior experiences, disposition, and environmental factors influence the way in which the individual interacts with and perceives the world. My goal is for machines to be capable of integrating this information to better understand individual end-users so as to optimize the machine-end-user relationship in a variety of different domains and settings.

Appendices

APPENDIX A
MUTUAL INFORMATION DRIVEN META-LEARNING FROM DEMONSTRATION

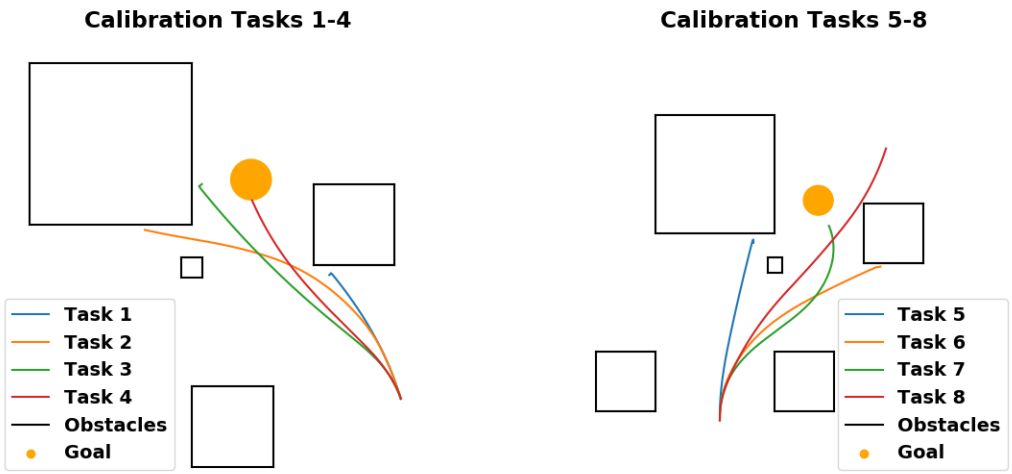
A.1 Calibration Tasks

We created a set of sixteen calibration tasks, which are depicted in Figure A.1. The tasks are "Wizard-of-Oz" [26] style rollouts representative of an agent's behavior. The rollouts were prerecorded to be consistent across participants, but participants believed an agent was driving the car. For each task, the car drives along the prerecorded trajectory and the participants provide corrective feedback. We calculated ground truth for each of the calibration tasks and use the ground truths and participant corrective feedback to train MIND MELD's parameters.

In our pilot study, we had a set of twelve calibration tasks; however, after running pilot participants, we concluded that this set of tasks did not adequately capture enough of the distribution of the task domain. The original calibration tasks did not have many obstacles, whereas the test task has participants navigating around multiple obstacles. Additionally, none of the trajectories successfully made it to the goal. In our final set of calibration tasks (Figure A.1), there are successful and unsuccessful trajectories as well as tasks with more obstacles that are more representative of the test task. We thought a wider and more representative set of tasks would more effectively capture the range of stylistic tendencies.

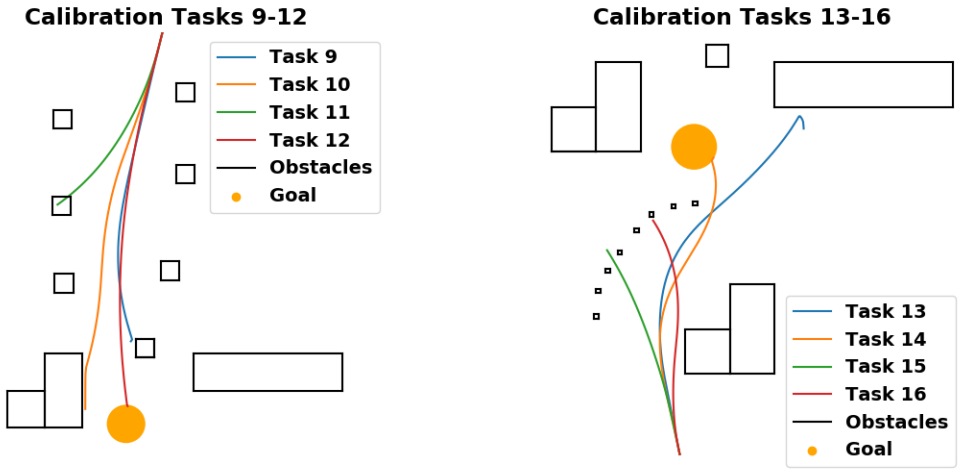
A.1.1 Calculating Ground Truth

For each prerecorded rollout in the calibration tasks, we calculated the optimal action to get to the goal. At each timestep along the rollout, we used Rapidly-exploring Random Trees (RRT)* [111] to find a path to the goal. RRT* treated the car as a point mass and expanded



(a) Calibration tasks 1-4

(b) Calibration tasks 5-8.

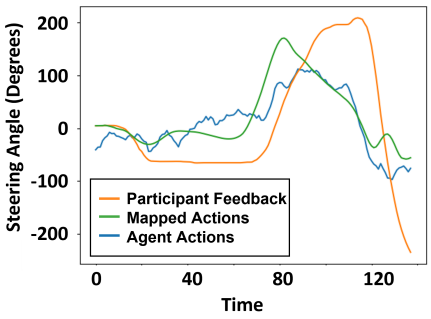


(c) Calibration tasks 9-12.

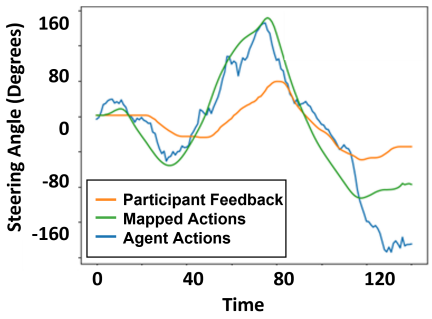
(d) Calibration tasks 13-16.

Figure A.1: This figure shows a graphical representation of the calibration tasks. The orange circle is the goal and the black squares are the obstacles in the environment. Each set of obstacles had four tasks with varying degrees of success. Calibration tasks 1-4 are not used to train MIND MELD and are used as practice to reduce novelty effects.

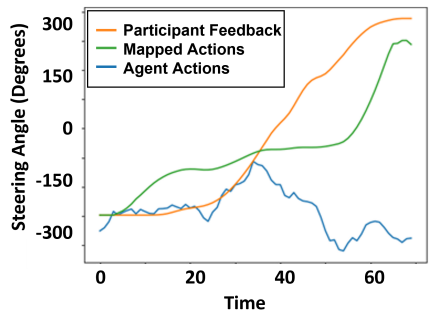
the obstacles accordingly, so that the RRT* path would not hit obstacles. We then employed an MPC controller [106], using an approximate dynamical model of the car to find the best steering angle to follow the RRT* path to the goal.



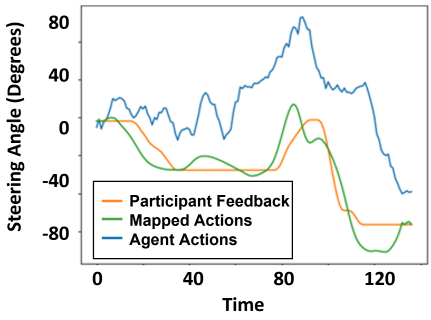
(a) This participant tended to over-correct by 34 degrees on average and provide delayed feedback by .86 seconds on average.



(b) This participant tended to under-correct by 32 degrees on average and provide delayed feedback by 1.7 seconds on average.



(c) This participant tended to over-correct by 127.2 degrees on average and provide anticipatory feedback by 1.0 seconds on average.



(d) This participant tended to neither over-correct nor under-correct and provided slightly anticipatory feedback by .14 seconds on average.

Figure A.2: This figure shows examples of suboptimal participant feedback and the corrections made by MIND MELD.

A.2 Additional Results

A.2.1 Path Efficiency

Although we do not explicitly tell participants to take the shortest path to the goal, we predict that participants typically attempt to optimize for path length to reduce the time required of them. However, participants tend to over-correct when providing feedback, resulting in DAgger taking a longer route to the goal. Because MIND MELD corrects for suboptimality and produces feedback more aligned with the intentions of the participant, we hypothesize that MIND MELD will achieve a shorter path to the goal compared to DAgger. To test

this hypothesis, we conduct a Kruskal-Wallis test comparing the path efficiency of MIND MELD and DAgger. We find that MIND MELD achieves a statistically significantly more efficient path to the goal ($H(1) = 14.33, p < .001$) compared to DAgger. We note that, because we do not explicitly ask participants to demonstrate the shortest path to the goal, our results are only speculative and require further investigation in future work.

A.2.2 Wheel Rate of Change

In this section, we investigate if the rate of change of the steering wheel during the calibration tasks correlates with the learned embeddings. We hypothesize that this relationship might exist because the rate at which the participants turn the wheel may be related to their suboptimal tendencies. If the rate of change differs greatly from the rate of change of the ground truths, then one might infer that participants are behaving in a suboptimal manner, which we would expect to be captured by the learned embeddings.

To investigate this hypothesis, we conduct a correlation analysis between the learned embeddings and the average rate of change of the wheel. We find that the rate of change significantly correlates with the learned embeddings ($r(116) = -.42, p < .001$), suggesting that our learned embeddings do capture information about the participants' rate of change of the wheel.

A.2.3 Examples of MIND MELD's ability to map suboptimal feedback

In Figure A.2, we show several examples of MIND MELD's ability to correct for heterogeneous and suboptimal feedback. Figure A.2(a) shows a participant who tended to provide over-corrected and delayed feedback. MIND MELD reduces the magnitude of the participant's actions and shifts the signal backward in time. In Figure A.2(b), the participant under-corrects and provides delayed feedback. As a result, MIND MELD increases the magnitude of the participant's feedback and shifts the feedback backwards in time. Fig Figure A.2(c) illustrates a participant who provides both anticipatory and over-corrected

feedback. MIND MELD shifts the feedback forward in time and reduces the magnitude of the feedback. Fig Figure A.2(d) shows a participant who neither over-corrected or under-corrected and provided only slightly anticipatory feedback. Due to the participant's feedback being near optimal, MIND MELD changes the feedback very little and instead reproduces the participant's original feedback.

A.2.4 Trust Results

The participants filled out a trust survey [115] after witnessing the agents' performance for each trial. We sought to compare the results from the trust surveys and the agent's performance to determine if there was a correlation between trust and performance. However, a repeated measures linear model for trust and distance to the goal did not pass Shapiro-Wilk's normality test. As such, we decided to use only the final trust survey in our analysis because it captured the participants' overall trust in the agent.

A.2.5 Workload Results

The NASA TLX [116] measures workload along six dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. After finding that participants rated MIND MELD as significantly lower workload than DAgger ($p = .005$) and BC ($p < .001$), we investigated whether the dimension of performance in the NASA TLX was causing the difference in workload scores. We removed the performance dimension from the calculation for workload and reanalyzed workload sans performance. The linear mixed-effects model with "workload sans performance" as the dependent variable and agent as the independent variable passed Shapiro-Wilk's test for normality ($p = .061$) and Levene's test for homoscedasticity ($p = .37$). An ANOVA with a Tukey post hoc found that "workload without performance" was still significantly lower for MIND MELD compared to DAgger ($p = .005$) and BC ($p < .001$) (Omnibus statistic: $F(2, 82) = 8.6$, $p < .001$).

Then we hypothesized that, since MIND MELD performed better and reached the goal more often, participants might have found MIND MELD less frustrating than DAgger and BC. To test this hypothesis, we removed the frustration dimension in addition to performance from the workload calculation. The linear mixed-effects model with "workload sans performance and frustration" as the dependent variable and agent as the independent variable passed Shapiro-Wilk's test for normality ($p = .21$) and Levene's test for homoscedasticity ($p = .45$). Surprisingly, an ANOVA with a Tukey post hoc found that "workload sans performance and frustration" was also significantly lower for MIND MELD than DAgger ($p = .006$) and BC ($p < .001$) (Omnibus statistic: $F(2, 82) = 8.1$, $p < .001$).

Therefore, we conclude that MIND MELD's lower workload ratings were not solely based on the agent's performance or frustration. Participants found teaching MIND MELD to be less mentally, physically, and temporally demanding and less effort than DAgger and BC.

A.3 Parametric Test Assumptions and Omnibus Results

To determine the proper statistical test to apply to our data, we first determine whether the residuals are normally distributed and homoscedastic. To do so, we apply Shapiro-Wilk's test for normality and Levene's test for homoscedasticity. In Table A.1, we report the results from Shapiro-Wilk's and Levene's test for our data and, based on the results, we applied a parametric or non-parametric test.

In Table A.2 we report the omnibus results for our subjective metrics and distance metric, the test employed, and the test statistics. If our data met parametric test assumptions, we applied an ANOVA, otherwise we applied Friedman's test.

A.4 Covariate Analysis

Each participant filled out surveys prior to starting the experiment to mitigate confounds related to demographics and prior experience. Through surveys, we collected information

Table A.1: We report our test results to determine if our data meets parametric test assumptions for each of our metrics. Based on the results, we specify the statistical test we applied.

Metric	Normality	Homoscedasticity	Test
Workload	$p = .03$	$p = .76$	Friedman
Likeability	$p = .03$	$p = .54$	Friedman
Intelligence	$p = .18$	$p = .82$	ANOVA
Trust	$p = .81$	$p = .09$	ANOVA
Distance	$p = .26$	$p = .31$	ANOVA
Video Game Experience	$p = .17$	$p = .002$	Spearman
Over-/Under-Correcting	$p = .24$	$p = .17$	Pearson
Anticipatory/ Delayed	$p = .57$	$p = .71$	Pearson
Wheel Rate	$p = .52$	$p = .71$	Pearson
Path Efficiency	$p < .001$	$p = .71$	Kruskal-Wallis

Table A.2: This table shows the omnibus test results.

Metric	Test	Statistics
Likeability	Friedman	$\chi^2(2) = 14.3, p < .001$
Intelligence	ANOVA	$F(2, 82) = 10.1, p < .001$
Trust	ANOVA	$F(2, 82) = 19.7, p < .001$
Workload	Friedman	$\chi^2(2) = 12.2, p = .002$
Distance	ANOVA	$F(2, 82) = 24.57, p < .001$

about participants' age, gender, personality traits, trust in automation, experience driving a car, experience playing video games, and experience driving a virtual car (surveys detailed in Appendix Subsection A.6.1). We ran a linear mixed-effects model for each dependent variable (workload, likeability, perceived intelligence, trust, and average distance) with the agent as the independent variable. For each linear model, we tested whether any of these covariates were significant or confounding factors using an ANOVA. We systematically added each covariate to the model based on whether it was significant or lowered the model's AIC. Overall, we did not find any covariates to be confounding factors in our analysis.

For the linear model with trust as the dependent variable, none of the covariates lowered

the AIC and the ANOVA did not show significance for any covariate. Adding gender lowered the AIC for the linear model with distance as the dependent variable; however, an ANOVA found gender to be insignificant. For the linear model with perceived intelligence as the dependent variable, no covariates lowered the AIC, but the ANOVA found that age was a significant covariate ($F(1, 82) = 4.5, p = .035$) with an effect in the negative direction. Adding age to the model did not decrease the significance of the independent variable.

Additionally, we tested whether there was an ordering effect for any of our significant dependent variables. We treated the agent's order as a categorical variable and added it as a covariate in each linear model. We did not find agent's order to be significant for any of the ANOVAs.

Note: Because the linear models for workload and likeability did not pass the normality test, we employed a Friedman's test to assess workload and likeability. A Friedman's test does not allow for covariates, so we could not conduct this analysis for workload and likeability.

A.5 Factor Analysis

We did not find any verified scales in prior work that measured prior experience in driving virtual or physical cars or playing video games, so we developed our own prior experience scales for driving a car, video games, and driving a virtual car. Based on guidance in [207], we generated an analogous set of eight items to measure each construct with minor differences based on the construct (items shown in Appendix Section Subsection A.6.1). We conducted a factor analysis [207] and found that the items loaded on two factors for *experience with driving* and *experience with video games*. Based on this factor analysis, we define two subscales which comprise *experience with driving* and *experience with video games*: confidence and familiarity. We sum both of these subscales to form one scale for experience. The factor analysis for *experience driving a virtual car* did not show that two factors were sufficient, so we did not use this scale in our results, as breaking the scale

into more than two factors would results in an inadequate number of items per subscale [113]. Additionally, we tested for internal consistency and we report Cronbach's alpha for each scale Subsection A.6.1. The subscales are shown below; we replaced *the activity* with *driving a car* or *playing video games* as appropriate.

Confidence Subscale

- I am comfortable with *the activity*.
- I do not feel comfortable with *the activity*.
- I think I am competent at *the activity*.
- I do not perform well at *the activity*.

Familiarity Subscale

- I *do the activity* on a regular basis.
- I am more experienced *at the activity* than the average person.
- I am not very experienced with *the activity*.
- I have not spent a lot of time *doing the activity*.

A.6 Surveys

A.6.1 Pre-Surveys

Big Five Personality Survey - Although we did not find significant results related to this survey, we collect information about the participants' personality via the Mini-IPIP questionnaire [134] to determine if personality correlates with the individual's learned embedding.

Agreeableness ($\alpha = .68$): Please rate the level of agreement you feel towards each of the given items on the scale from 1 (strongly disagree) to 5 (strongly agree).

- I sympathize with others' feelings.

- I feel others' emotions.
- I am not really interested in others.
- I am not interested in other people's problems.

Neuroticism ($\alpha = .67$) **[134]:** Please rate the level of agreement you feel towards each of the given items on the scale from 1 (strongly disagree) to 5 (strongly agree).

- I have frequent mood swings.
- I get upset easily.
- I am relaxed most of the time.
- I seldom feel blue.

Conscientiousness ($\alpha = .75$) **[134]:** Please rate the level of agreement you feel towards each of the given items on the scale from 1 (strongly disagree) to 5 (strongly agree).

- I get chores done right away.
- I like order.
- I often forget to put things back in their proper place.
- I make a mess of things.

Openness ($\alpha = .70$) **[134]:** Please rate the level of agreement you feel towards each of the given items on the scale from 1 (strongly disagree) to 5 (strongly agree).

- I have a vivid imagination.
- I have difficulty understanding abstract ideas.
- I am not interested in abstract ideas.

- I do not have a good imagination.

Extraversion ($\alpha = .82$) [134]: Please rate the level of agreement you feel towards each of the given items on the scale from 1 (strongly disagree) to 5 (strongly agree).

- I am the life of the party.
- I talk to a lot of different people at parties.
- I don't talk a lot.
- I keep in the background.

Experience Survey - We measured participants' prior experience with driving cars, playing video games, and driving virtual cars to determine if their prior experience correlated with an individual's learned embedding or the agent's performance.

Experience Driving Car ($\alpha = .93$): Please rate the level of agreement you feel towards each of the given items on the scale from 1 (strongly disagree) to 5 (strongly agree).

- I am comfortable driving a car.
- I drive a car on a daily basis.
- I am more experienced driving a car than the average person.
- I am not very experienced driving a car.
- I do not feel comfortable driving a car.
- I think I am competent at driving a car.
- I have not spent a lot of time driving cars.
- I do not perform well at driving cars.

Experience Playing Videos Games ($\alpha = .93$)**:** Please rate the level of agreement you feel towards each of the given items on the scale from 1 (strongly disagree) to 5 (strongly agree).

- I am comfortable playing video games.
- I play video games on a regular basis.
- I am more experienced playing video games than the average person.
- I am not very experienced with playing video games.
- I do not feel comfortable playing video games.
- I think I am competent at playing video games.
- I have not spent a lot of time playing playing video games.
- I do not perform well at video games.

Experience Driving a Virtual Car ($\alpha = .86$)**:** Please rate the level of agreement you feel towards each of the given items on the scale from 1 (strongly disagree) to 5 (strongly agree).

- I am comfortable driving a virtual car.
- I play racing games on a regular basis.
- I am more experienced driving virtual cars than the average person.
- I am not very experienced with driving a virtual car.
- I do not feel comfortable driving a virtual car.
- I think I am competent at driving a virtual car.
- I am not very experienced with playing video games that involve race cars.

- I do not perform well at video games involving race cars.

Pre-Trust Survey - We measure the participants' trust in automation via the survey presented in [137] to measure the participants' dispositional trust and determine if trust in automation impacted how a participant rated the agent.

Trust in Automation ($\alpha = .71$) [137]: Please rate the level of agreement you feel towards each of the given items on the scale from 1 (strongly disagree) to 5 (strongly agree).

- I think that automated devices used in medicine, such as CT scans and ultrasound, provide very reliable medical diagnosis.
- Automated devices in medicine save time and money in the diagnosis and treatment of disease.
- If I need to have a tumor in my body removed, I would choose to undergo computer-aided surgery using laser technology because it is more reliable and safer then manual surgery.
- Automated systems used in modern aircraft, such as automatic landing systems, have made air journeys safer.
- ATMs provide a safeguard against the inappropriate use of an individual's bank account by dishonest people.
- Automated devices used in aviation and banking have made work easier for both employees and customers.
- Even though the automatic cruise control in my car is set at a speed below the speed limit, I worry when I pass a police radar speed trap in case the automatic control is not working properly.
- Manually sorting through card catalogues is more reliable than computer-aided searches for finding items in a library.

- I would rather purchase an item using a computer than have to deal with a sales representative on the phone because my order is more likely to be correct using the computer.
- Bank transactions have become safer with the introduction of computer technology for the transfer of funds.
- I feel safer depositing my money at an ATM than with a human teller.
- I have to tape an important TV program for a class assignment. To ensure that the correct program is recorded, I would use the automatic programming facility on my VCR rather than manual taping.

A.6.2 Post-Surveys

Trust ($\alpha = .96$) [115]: Please rate the level of agreement you feel towards each of the given items on the scale from 1 (strongly disagree) to 5 (strongly agree).

- The system is deceptive .
- The system behaves in an underhanded manner.
- I am suspicious of the system's intent, action, or outputs.
- I am wary of the system.
- The system's actions will have a harmful or injurious outcome.
- I am confident in the system.
- The system provides security.
- The system has integrity.
- The system is dependable.

- The system is reliable.
- I can trust the system.
- I am familiar with the system.

Likeability ($\alpha = .95$) **[117]:** Please choose the point which best describes your feeling or your impression of the agent on the scale from 1 to 9.

- Awful - Nice
- Unfriendly - Friendly
- Unkind - Kind
- Unpleasant - Pleasant

Perceived Intelligence ($\alpha = .95$) **[117]:** Please choose the point which best describes your feeling or your impression of the agent on the scale from 1 to 9.

- Incompetent - Competent
- Ignorant - Knowledgeable
- Irresponsible - Responsible
- Unintelligent - Intelligent
- Foolish - Sensible

Workload [116] Please rate the task based on the following factors on a scale from 0-100.

- Mental Demand - How much mental activity was required to complete the task?
- Physical Demand - How much physical activity was required to complete the task?

- Temporal Demand - How much time pressure did you feel while completing the task?
- Performance - How well did you think you completed the task?
- Effort - How hard did you have to work to complete the task?
- Frustration - How did you feel during the task?

The subject is then presented with pairwise comparisons for each measure and told to: Please choose the scale title that represents the most important contributor to workload for the specific task(s) you performed in this experiment.

APPENDIX B

PERSONALIZED TEACHING VIA RECIPROCAL MUTUAL INFORMATION DRIVEN META-LEARNING FROM DEMONSTRATION

B.1 Reciprocal MIND MELD Architecture

Figure B.1 shows the steps in the Reciprocal MIND MELD framework. To determine the robotic feedback that should be provided to the demonstrator, we first learn a semantically meaningful embedding space. The robot then provides feedback to the demonstrator based upon the distance from the perfect embedding in each semantically meaningful dimension. For example, the robot provides feedback in the over-/under-correcting dimension based on the distance, $\epsilon_{o/u}$. We then re-estimate the embedding after robotic feedback. In Study 1 and Study 2, participants experience four and five rounds of robotic feedback respectively. Between rounds, if the participant improves their feedback but is still not within the first quartile, the robot says, "That is better but..." followed by the appropriate feedback as shown in Table B.1.

Table B.1 shows the feedback provided to the demonstrator in Study 1 for the over-/under-correcting dimension. If a demonstrator is in a quartile that is farther from the perfect demonstrator, the feedback is intended to shift their embedding by a larger amount than demonstrators in quartiles closer to the perfect demonstrator. Analogous feedback is provided for the anticipatory/delayed dimension in Study 2 and Study 3. In all conditions, regardless of whether feedback is provided, the robot interacts with the participant and says "Please provide me with a demonstration" before each round.

Figure B.2 illustrates the embedding space in which the size of the points represents the magnitude by which a participant is anticipatory/delayed. Q1-Q4 indicate the quartiles for anticipatory/delayed. Points that are farther from the decision boundaries represent

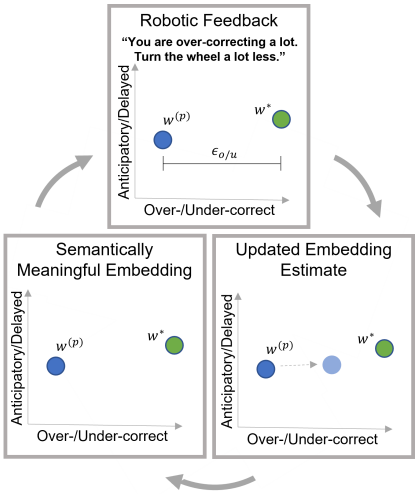


Figure B.1: This figure shows our Reciprocal MIND MELD framework. $\epsilon_{o/u}$ is the distance between the participant's current embedding, $w^{(p)}$, and the perfect embedding, w^*, along the over-/under-correcting dimension.

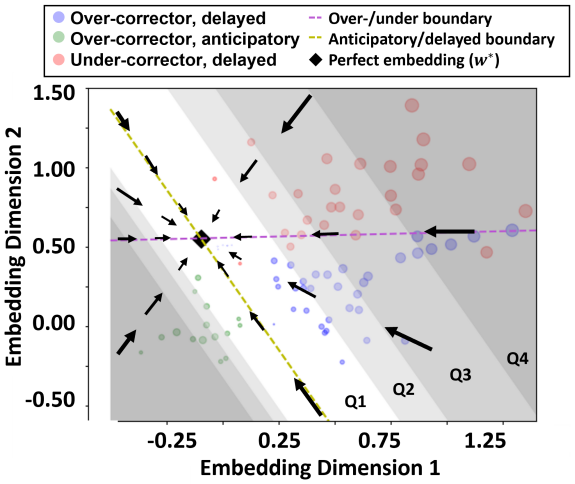


Figure B.2: This figure depicts the learned embedding space and decision boundaries. Each point represents the embedding of a demonstrator, and the diameter represents the magnitude by which participants are anticipatory/delayed. Q1-Q4 indicate quartiles one through four for the anticipatory/delayed dimension.

participants who are more suboptimal in their demonstrations. The objective is to provide robotic feedback to shift a participants embedding into Q1.

Table B.1: This table shows the feedback a participant receives based on their quartile and study condition for Study 1. Analogous feedback for the Cooperative condition is provided in Study 2 for the anticipatory/delayed dimension in addition to the over-/under-correcting dimension.

Cooperative Quartile	**Adversarial Quartile**	**Robotic Feedback**
First	Fourth	"Your feedback is good! Keep it up."
Second	Third	"You are slightly over-/under-correcting. Please turn the wheel a bit more/less."
Third	Second	"You are over-/under-correcting. Please turn the wheel more/less."
Fourth	First	"You are over-/under-correcting a lot. Please turn the wheel a lot more/less."

B.2 MIND MELD Architecture

Below we describe the MIND MELD architecture and discuss the alterations to learn a semantically meaningful embedding space.

B.2.1 Network Architecture

Figure B.3 shows the MIND MELD architecture. The three main components of the architecture are: 1) the bi-directional long short-term memory (LSTM) encoder, $\mathcal{E}_{\phi'}: A \rightarrow Z$, 2) the prediction subnetwork, $f_\theta: Z \times W \rightarrow \mathbb{R}$, and 3) the mutual information subnetwork, $q_\phi: Z \times \mathbb{R} \rightarrow \mathcal{N}_W$. Our goal is to improve upon the corrective feedback, $a_t^{(p)}$, from a demonstrator, p. The corrective feedback from the human demonstrator from $t - \Delta t : t + \Delta t$ is fed into the bi-directional LSTM, $\mathcal{E}_{\phi'}$, to extract an encoding, $z^{(p)}_{t-\Delta t:t+\Delta t}$. The f_θ subnetwork takes in the encoding, $z^{(p)}_{t-\Delta t:t+\Delta t}$, and the personalized embedding, $w^{(p)}$, and learns the predicted difference, $\hat{d}_t^{(p)}$, between the optimal label, o_t, and the human's corrective label, $a_t^{(p)}$. The q_ϕ subnetwork learns to map the difference, $\hat{d}_t^{(p)}$, and the encoding, $z^{(p)}_{t-\Delta t:t+\Delta t}$, to a posterior distribution over the person's embedding, $w^{(p)}$. We estimate an individual's learned embedding, $\hat{w}^{(p)}$, by sampling from the approximate posterior [104]. $w^{(p)}$ is initialized based upon the prior, $\hat{w}^{(p)} \sim \mathcal{N}(0, 1)$.

B.2.2 Loss Function for Semantic Meaning

To learn a semantically meaningful embedding space, we add an additional network head, p_ψ, to the MIND MELD architecture to aid in learning the embedding space. p_ψ is a linear layer to encourage the embedding space to be linearly separable. We utilize a mean squared error (MSE) loss, $l = \frac{1}{N} \sum_i \left(p_\psi(w^{(i)}) - m^{(i)}_{o/u,a/d} \right)^2$, to train the network to predict the suboptimal tendencies, $m_{o/u}$ and $m_{a/d}$, (i.e., the magnitude by which a demonstrator over-/under-corrects and is delayed/anticipatory) given the personalized embedding. We calculate $m_{o/u}$ and $m_{a/d}$ via dynamic time warping (DTW) [114] between the demonstrations and

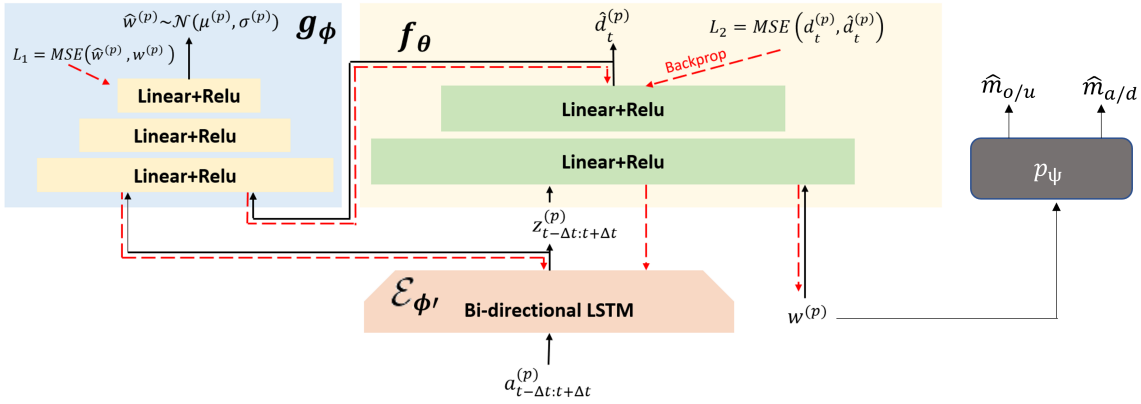


Figure B.3: This figure shows the MIND MELD network architecture. The inputs to the architecture are a demonstrator, p's, corrective labels, $a^{(p)}_{(t-\Delta t:t+\Delta t)}$, from time $t - \Delta t$ to $t + \Delta t$ and the personalized embedding, $w^{(p)}$. The bi-directional LSTM extracts sequential information about the demonstrator's feedback. The f_θ subnetwork learns the predicted difference, $\hat{d}_t^{(p)}$, by minimizing the mean squared error (MSE) between $\hat{d}_t^{(p)}$ and the true difference, $d_t^{(p)} = a_t^{(p)} - o_t$, between the demonstrator's corrective feedback, $a_t^{(p)}$, and the optimal label, o_t. The re-creation subnetwork q_ϕ maximizes mutual information between the personalized embedding, $w^{(p)}$, the encoding $z^{(p)}_{(t-\Delta t:t+\Delta t)}$, and the learned difference, $\hat{d}_t^{(p)}$ to estimate the learned embedding, $\hat{w}^{(p)}$ [1, 2]. We add the additional network head, p_ψ, to learn a semantically meaningful embedding space. The outputs $\hat{m}_{o/u}$ and $\hat{m}_{a/d}$ are estimates for how much a demonstrator is over-/under-correcting and anticipatory/delayed.

the optimal labels in the calibration tasks. This loss helps to ensure that our embedding space can be translated into actionable robotic feedback, (i.e., the magnitude by which a demonstrator over-/under-corrects and is delayed/anticipatory) given the personalized embedding.

B.2.3 Variational Inference

We assume that humans provide heterogeneous and distinct styles when providing corrective feedback to the robot. A person's corrective style is encapsulated in the embedding, $w^{(p)}$, for person, p. To learn $w^{(p)}$, we maximize the lower bound on the mutual information between the learned embedding, $w^{(p)}$, and the predicted difference between the human feedback and the optimal feedback, $\hat{d}_t^{(p)}$ (Equation B.1). Intuitively, maximizing mutual information

means that observing the difference, $\hat{d}_t^{(p)}$, will reduce uncertainty about the personalized embedding.

In Equation B.1, the mutual information between $z^{(p)}$, $\hat{d}_t^{(p)}$, and personalized embedding, $w^{(p)}$, is denoted as $I(w^{(p)}; z^{(p)}, \hat{d}_t^{(p)})$. However, maximizing the mutual information requires access to an intractable posterior distribution, $P(w^{(p)}|z^{(p)}, \hat{d}_t^{(p)})$; therefore, we employ variational inference and a lower bound on mutual information to estimate the distribution using q_ϕ [108]. The variational lower bound is $L_I(f_{\theta|w}, q_{\phi|\theta})$.

$$I(w^{(p)}; z^{(p)}, \hat{d}_t^{(p)}) = H(w^{(p)}) - H(w^{(p)}|z^{(p)}, \hat{d}_t^{(p)}) \geq \tag{B.1}$$
$$\mathbb{E}[log(q_\phi(w^{(p)}|z^{(p)}, \hat{d}_t^{(p)}))] + H(w^{(p)}) = L_I(f_{\theta|w}, q_{\phi|\theta})$$

The MIND MELD architecture utilizes two loss functions, one to learn the personalized embedding, $w^{(p)}$, and another to learn the amount by which a person's feedback is suboptimal, $\hat{d}_t^{(p)}$, as shown in Figure B.3. For the q_ϕ subnetwork, we minimize the mean squared error between the sampled approximation of the embedding, $\hat{w}^{(p)}$, and the personalized embedding, $w^{(p)}$, which is equivalent to maximizing the log-likelihood of the posterior. The loss function for the f_θ subnetwork is the mean squared error between the predicted difference, $\hat{d}_t^{(p)}$, and the difference between the human feedback and the optimal labels, $d_t^{(p)} = a_t^{(p)} - o_t$. These two losses are summed (Equation B.2) and backpropagated through the layers and the input embedding, $w^{(p)}$, so that the embedding converges to reflect a person's feedback style. At test time, the MIND MELD network parameters θ, ϕ, and ϕ' are frozen. We then backpropagate only through $w^{(p)}$, to learn an embedding that encapsulates a participant's suboptimal style.

$$L_{(\theta,\phi,\phi',w)} = L_{1_{(\theta,\phi,\phi')}} + \lambda L_{2_{(\theta,\phi')}} \tag{B.2}$$

$$L_{1_{(\theta,\phi,\phi')}} = \frac{1}{K+1}\sum_{k=0}^{K} ||\hat{w}_k^{(p)} - w_k^{(p)}|| \tag{B.3}$$

$$L_{2_{(\theta,\phi')}} = ||d_k^{(p)} - \hat{d}_k^{(p)}|| \tag{B.4}$$

B.3 Calibration and Novel Tasks

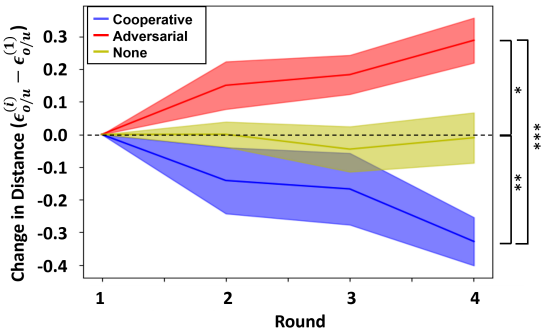
Figure B.6 a-d shows the calibration tasks employed in the study. These tasks of pre-recorded policy rollouts and are consistent for all participants. Figure B.6 e-g shows the novel tasks in which participants provide demonstrations to teach the car to get from the start location to the goal.

B.4 Additional Results from Study 1

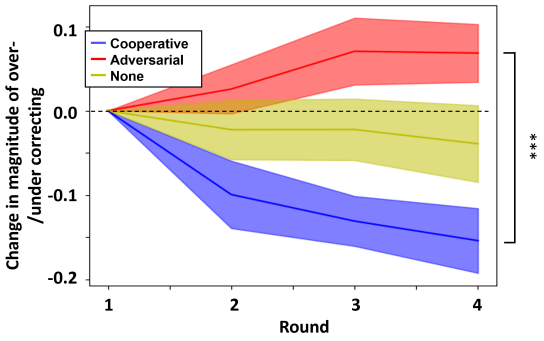
Table B.2: This table shows the mean, (standard deviation), and test statistics of the subjective metrics and $\Delta\epsilon_{o/u}$ for Study 1. Δ Trust and Δ Fluency describe the change in Trust and Fluency respectively between rounds one and four.

	Cooperative	**Adversarial**	**None**	Test Statistic	p-value
$\Delta\epsilon_{o/u}$	0.33 (0.2)	-0.30 (0.2)	0.01 (0.2)	$F(2,24) = 20.2$	$p < .001$
Workload	37.5 (16.4)	46.1 (19.5)	53.5 (11.6)	$F(2,24) = 2.21$	$p = .132$
Likeability	6.69 (2.0)	6.81 (1.5)	6.86 (1.4)	$F(2,24) = .024$	$p = .978$
Intelligence	6.31 (1.6)	5.57 (1.1)	6.24 (1.4)	$F(2,24) = 1.03$	$p = .372$
Δ Trust	0.56 (0.4)	-0.01 (0.4)	0.05 (0.2)	$F(2,24) = 5.15$	$p = .014$
Δ Fluency	0.34 (0.4)	-0.13 (0.3)	-0.04 (0.4)	$F(2,24) = 5.10$	$p = .014$

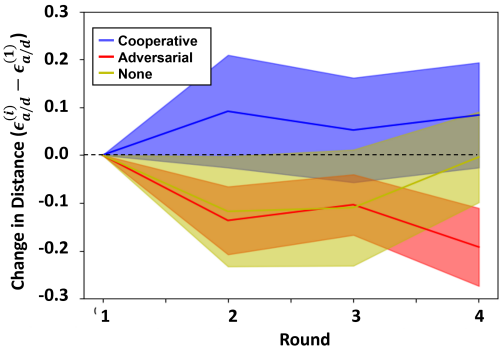
Figure B.4(a) shows the change in the distance ($\epsilon_{o/u}^{(4)} - \epsilon_{o/u}^{(1)}$) in the over-/under-correcting dimension between round one and rounds one through four. Figure B.4(b) shows the change between rounds one and four in the amount by which the participant over-/under-corrects as calculated via dynamic time warping (DTW) between the participant demonstrations and the



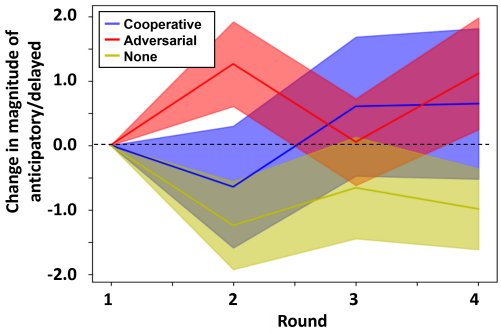
(a) Difference between the embedding distance at each round, $\epsilon_{o/u}$, and the embedding distance at round one, $\epsilon^{(1)}_{o/u}$.



(b) Amount by which participants over-/under-correct at each round minus the amount by which a participant over-/under-corrects in round one as calculated by dynamic time warping.



(c) Difference between the embedding distance at each round, $\epsilon_{a/d}$, and the embedding distance at round one, $\epsilon^{(1)}_{a/d}$.



(d) Amount by which participants are anticipatory/delayed at each round minus the amount by which a participant is anticipatory/delayed in round one as calculated by dynamic time warping.

Figure B.4: This figure shows the average distance of the embedding and the dynamic time warping results for the over-/under-correcting dimension and anticipatory/delayed dimension for Study 1.

optimal labels. The similarity in trends between Figure B.4(a) and Figure B.4(b) suggests that robotic feedback is not only able to shift a participant's embedding but robotic feedback is also able to alter the amount by which a participant over-/under-corrects. This finding lends support to the idea that the distance from the embedding to the decision boundary is a good measure of how much a participant over-/under-corrects.

Because data does not meet parametric assumptions, we apply Friedman's test to determine if there is a statistically significant difference in how much an individual over-/under-corrects in round one versus round four as determined via DTW. We find that participants

over-/under-correct significantly less in round four compared to round one in the Cooperative condition ($\chi^2(1) = 9.00$, $p = .003$). We find that the opposite is true in the Adversarial condition, with participants over-/under-correcting more in round four versus round one ($\chi^2(1) = 5.44$, $p = .020$). We do not find a significant difference for the None condition ($\chi^2(1) = .111$, $p = .740$).

We additionally compare the DTW results in round four between conditions. We find significance in an omnibus ANOVA test ($F(2, 24) = 8.99$, $p = .001$). We applied Tukey post-hoc test and find that the Cooperative agent results in the participant significantly over-/under-correcting less compared to the Adversarial condition ($p < .001$). These results suggest that a participant provides demonstrations closer to the optimal by the fourth round in the Cooperative agent condition.

Figure B.4(c) and Figure B.4(d) show the change in the amount by which a participant provides anticipatory/delayed feedback as calculated by the distance from the perfect demonstrator in embedding space and DTW respectively. We show in Figure B.4(c) that as participants improve in the over-/under-correcting dimension, they tend to become worse in the anticipatory/delayed dimension when no feedback is provided. This suggests that the task of improving participants demonstration quality in both the over-/under-correcting dimension and the anticipatory/delayed dimension may be particularly difficult since improving in the over-/under-correcting dimension tends to produce greater suboptimality in the anticipatory/delayed dimension.

Table B.2 shows the results of the subjective metrics. After each round, participants completed surveys measuring trust [115] and team fluency [122]. At the end of the study, participants completed surveys measuring workload [116] and likeability and perceived intelligence [117]. By applying a one-way ANOVA with Tukey post-hoc, we find that participants' trust increased significantly more ($F(2, 24) = 5.15$, $p = .014$) in Cooperative compared to Adversarial ($p = .020$) and None ($p = .038$). We do not find significance between Adversarial and None. Similar trends emerge for change in team fluency. We

find that participants report statistically significantly greater positive change in fluency ($F(2, 24) = 5.10$, $p = .014$) in Cooperative compared to Adversarial ($p = .017$) and close to significant change compared to None ($p = .052$). Again, we do not find significant difference between Adversarial and None.

While we do not find significance between conditions with regards to the other subjective metrics, we do note some trends that merit discussion. Surprisingly, we find that Cooperative is rated as requiring lower workload compared to Adversarial and None, despite participants likely having to exert similar or additional mental effort to comply with the demands of the robot. We also find that the Cooperative robot is rated as more intelligent compared to both the Adversarial and None teachers.

B.5 Additional Results from Study 2

Figure B.5 shows the embedding distance in the over-/under-correcting dimension (Figure B.5(a)) and the anticipatory/delayed dimension (Figure B.5(c)). We additionally show the results of DTW for over-/under-correcting (Figure B.5(b)) and anticipatory/delayed (Figure B.5(d)). We find that the embedding distance in the over-/under-correcting dimension as well as the amount that a participant over-/under-corrects as determined via DTW both decrease the most in the Simultaneous condition. We also find that participants improve in the anticipatory/delayed dimension the most in the Simultaneous condition compared to Greedy and None. Although the difference between Simultaneous and None is small, this finding is noteworthy because we found in Study 1 that participants tend to become considerably worse in the anticipatory/delayed dimension as they improve in the over-/under-correcting dimension. In this study, we show that with Simultaneous feedback, participants improve in both dimensions.

Table B.3 shows the mean and standard deviations of the change in the embedding distance as well as subjective metrics for each condition. Study 2 uses the same subjective metrics as Study 1: trust and team fluency after each round and workload, likeability, and

Table B.3: This table shows the mean, (standard deviation), and test statistics of the subjective metrics, $\Delta\epsilon_{o/u+a/d}$, $\Delta\epsilon_{o/u}$, and $\Delta\epsilon_{a/d}$ for Study 2. Δ Trust, Δ Fluency, and Δ Understanding describe the change in Trust, Fluency, and Understanding respectively between rounds one and five.

	Simultaneous	**Greedy**	**None**	F(2,36)	p-value
$\Delta\epsilon_{o/u+a/d}$	0.33 (0.25)	0.11 (0.31)	0.04 (0.28)	3.77	$p = .033$
$\Delta\epsilon_{o/u}$	0.267 (0.30)	0.09 (0.44)	-0.02 (0.27)	2.19	$p = .126$
$\Delta\epsilon_{a/d}$	0.07 (0.19)	0.06 (0.23)	0.02 (0.22)	.192	$p = .826$
Workload	50.9 (12.7)	51.3 (13.7)	43.9 (17.5)	1.05	$p = .360$
Likeability	6.88 (2.16)	7.5 (1.87)	6.58 (1.66)	.790	$p = .462$
Intelligence	7.34 (2.03)	6.75 (1.72)	5.71 (1.22)	3.10	$p = .057$
Δ **Trust**	0.54 (0.60)	0.37 (0.68)	-0.77 (0.46)	3.81	$p = .032$
Δ **Fluency**	0.78 (0.90)	0.25 (0.58)	-0.23 (0.47)	7.23	$p = .002$
Δ **Understanding**	0.61 (0.62)	0.15 (0.58)	0.14 (0.69)	2.33	$p = .112$

perceived intelligence at the end of the study. In Study 2, we also utilize the Robot Self-Efficacy Scale to measure a participant's level of understanding [240] after each round. For each metric, we employ an ANOVA comparing the three conditions: Simultaneous, Greedy, and None. If there is a significant main effect, then we conduct a Tukey post-hoc test. As stated in the main paper, we find that Simultaneous results in significantly increased trust ($p = .032$) and team fluency ($p = .002$) ratings. Although not significant, we find that participant's understanding of the robot increased more in the Simultaneous condition compared to None and Greedy. Also, participants perceived the Simultaneous feedback as more intelligent than Greedy or None.

B.6 Additional Results from Study 3

Table B.4 lists the mean and standard deviations of the change in the embedding distance and the subjective metrics between the Feedback and No Feedback conditions. Study

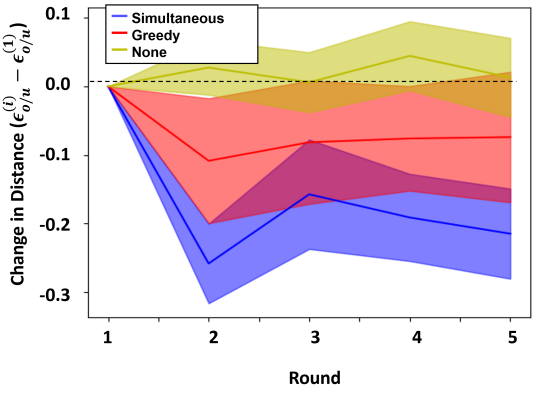
3 employed the same subjective metrics as Study 2: trust, team fluency, understanding, workload, likeability, and perceived intelligence. To compare between conditions, we utilized either a one-tailed t-test, if the model passed normality and homoscedasticity assumptions or a one-tailed Wilcoxon Signed Rank test, a non-parametric test. We employed one-tailed tests because we hypothesized that the Feedback condition would be better on all metrics (higher for change in embeddings, likeability, perceived intelligence, trust, fluency, and understanding and lower for workload) than No Feedback.

The amount that a person's embedding improved in the over-/under-correcting dimension, $\Delta\epsilon_{o/u}$, was significantly higher in the Feedback condition compared to No Feedback ($p = .006$). Although not significant, the amount that a person's embedding changed in the anticipatory/delayed, $\Delta\epsilon_{a/d}$, dimension was an improvement in the Feedback condition and got worse in the No Feedback condition. Additionally, the sum of these dimensions, $\Delta\epsilon_{o/u+a/d}$, was significantly improved in the Feedback condition compared to No Feedback ($p = .009$).

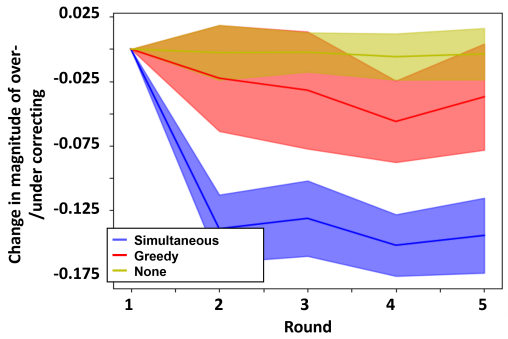
In terms of subjective metrics, we find the Feedback condition to be significantly lower in terms of workload compared the No Feedback condition ($p = .039$). Also, we find that participant's trust increased significantly more in the Feedback condition compared to No Feedback ($p = .019$). We additionally find that participants' perceived intelligence of the robot is trending towards being significantly higher for the Feedback condition compared to No Feedback ($p = .081$). Lastly, while not significant, participants' perceived team fluency and understanding increased more in the Feedback condition versus the No Feedback condition.

Table B.4: This table shows the mean, (standard deviation), and test statistics of the subjective metrics, $\Delta\epsilon_{o/u+a/d}$, $\Delta\epsilon_{o/u}$, and $\Delta\epsilon_{a/d}$ for Study 3. Δ Trust, Δ Fluency, and Δ Understanding describe the change in Trust, Fluency, and Understanding respectively between the first and last round.

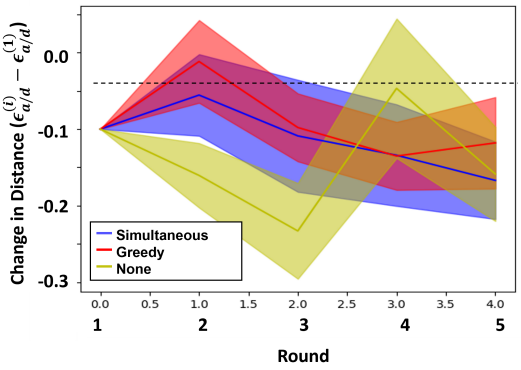
	Feedback	**No Feedback**	Test Statistic	p-value
$\Delta\epsilon_{o/u+a/d}$	0.17 (0.35)	-0.05 (0.37)	$t(57.8) = 2.45$	$p = .009$
$\Delta\epsilon_{o/u}$	0.16 (0.28)	0.004 (0.19)	$t(52.0) = 2.62$	$p = .006$
$\Delta\epsilon_{a/d}$	0.01 (0.36)	-0.05 (0.33)	$t(57.3) = .724$	$p = .236$
Workload	44.3 (15.5)	51.4 (15.3)	$t(58.0) = -1.79$	$p = .039$
Likeability	7.14 (1.67)	7.24 (1.65)	$t(58.0) = -.233$	$p = .592$
Intelligence	7.09 (1.14)	6.62 (1.40)	$t(55.7) = 1.42$	$p = .081$
Δ Trust	0.79 (0.75)	0.39 (0.51)	$Z = -2.07$	$p = .019$
Δ Fluency	0.49 (0.56)	0.31 (0.54)	$t(57.9) = 1.25$	$p = .108$
Δ Understanding	0.49 (0.53)	0.42 (0.86)	$Z = -.790$	$p = .430$



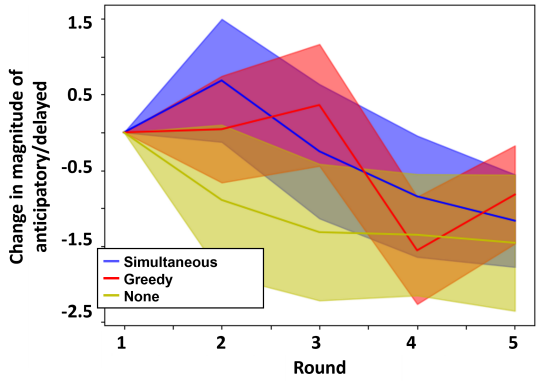
(a) Difference between the embedding distance at each round, $\epsilon_{o/u}$, and the embedding distance at round one, $\epsilon^{(1)}_{o/u}$.



(b) Amount by which participants over-/under-correct at each round minus the amount by which a participant over-/under corrects in round one as calculated by dynamic time warping.

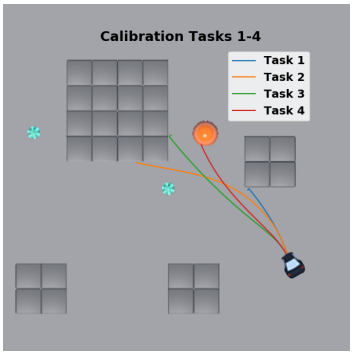


(c) Difference between the embedding distance at each round, $\epsilon_{a/d}$, and the embedding distance at round one, $\epsilon^{(1)}_{a/d}$.

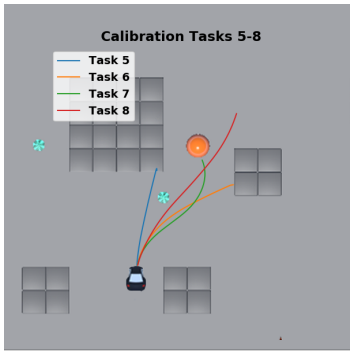


(d) Amount by which participants are anticipatory/delayed at each round minus the amount by which a participant is anticipatory/delayed in round one as calculated by dynamic time warping.

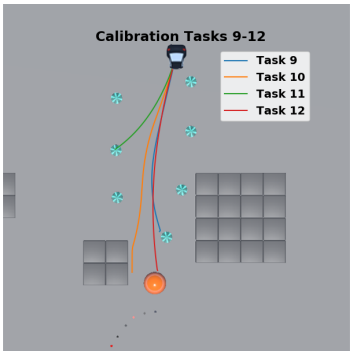
Figure B.5: This figure shows the average distance of the embedding and the dynamic time warping results for the over-/under-correcting dimension and anticipatory/delayed dimension for Study 2.



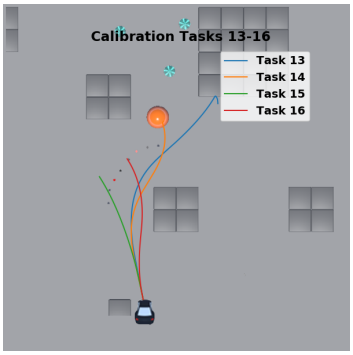
(a) Calibration tasks 1-4



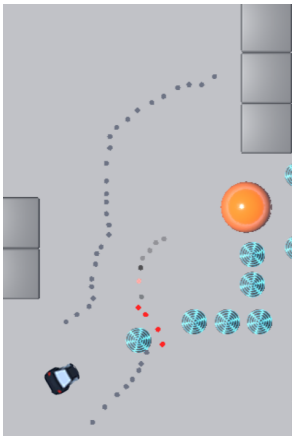
(b) Calibration Tasks 5-8



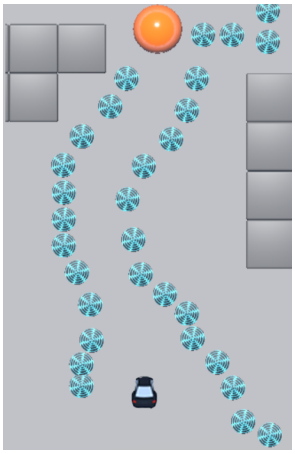
(c) Calibration Tasks 9-12



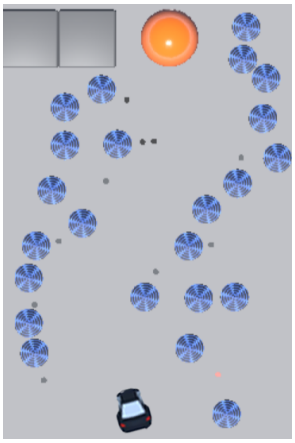
(d) Calibration Tasks 13-16



(e) Novel Task



(f) Novel Task



(g) Novel Task

Figure B.6: This figure depicts the calibration tasks and novel tasks in the study. Figure B.6(a)-Figure B.6(d) show the calibration tasks. The car is the starting location and the orange ball is the goal location. The rest of the objects in the scene are obstacles. Each line represents one of the pre-recorded rollouts, which are a mix of successful and unsuccessful trajectories. Figure B.6(e)-Figure B.6(g) show the environment for the novel tasks. There are no rollout lines because the trajectories were dependent on participant input. The calibration tasks are simpler, have less obstacles, and less turns than the novel tasks.

B.7 Model Assumptions (Table B.5)

Table B.5: This table lists the statistical models and tests utilized in our analysis. The dependent variable (DV) and independent variable (IV) are specified for each model. We tested for normality using the Shapiro-Wilk test. When the IV is categorical, we employed Levene's test for homoscedasticity, otherwise, we employed the Breusch-Pagan test. If the model did not pass normality or homoscedasticity, then we used a non-parametric version of the statistical test.

Study 1				
DV	IV	Test	Normality	Homoscedasticity
$\epsilon_{o/u}^{(i)}$	Cooperative, $i = 1, 4$	Friedman's	N/A	N/A
$\epsilon_{o/u}^{(i)}$	Adversarial, $i = 1, 4$	rANOVA	$p = .424$	$p = .149$
$\epsilon_{o/u}^{(i)}$	None, $i = 1, 4$	rANOVA	$p = .706$	$p = .856$
DTW Round i	Cooperative, $i = 1, 4$	Friedman's	N/A	N/A
DTW Round i	Adversarial, $i = 1, 4$	Friedman's	N/A	N/A
DTW Round i	None, $i = 1, 4$	Friedman's	N/A	N/A
$\Delta\epsilon_{o/u}$	Condition	ANOVA	$p = .547$	$p = .931$
DTW Last Round	Condition	ANOVA	$p = .179$	$p = .855$
Workload	Condition	ANOVA	$p = .598$	$p = .454$
Likeability	Condition	ANOVA	$p = .770$	$p = .459$
Intelligence	Condition	ANOVA	$p = .571$	$p = .632$
ΔTrust	Condition	ANOVA	$p = .907$	$p = .925$
ΔTeam Fluency	Condition	ANOVA	$p = .457$	$p = .558$
Study 2				
DV	IV	Test	Normality	Homoscedasticity
$\epsilon_{o/u+a/d}^{(i)}$	Simultaneous, $i = 1, 5$	rANOVA	$p = .092$	$p = .826$
$\epsilon_{o/u+a/d}^{(i)}$	Greedy, $i = 1, 5$	rANOVA	$p = .167$	$p = .723$
$\epsilon_{o/u+a/d}^{(i)}$	None, $i = 1, 5$	rANOVA	$p = .081$	$p = .194$
$\Delta\epsilon_{o/u+a/d}$	Condition	ANOVA	$p = .708$	$p = .614$
$\Delta\epsilon_{o/u}$	Condition	ANOVA	$p = .100$	$p = .448$
$\Delta\epsilon_{a/d}$	Condition	ANOVA	$p = .565$	$p = .589$
Workload	Condition	ANOVA	$p = .573$	$p = .373$
Likeability	Condition	ANOVA	$p = .752$	$p = .587$
Intelligence	Condition	ANOVA	$p = .238$	$p = .453$
ΔTrust	Condition	ANOVA	$p = .290$	$p = .368$
ΔTeam Fluency	Condition	ANOVA	$p = .091$	$p = .201$
ΔUnderstanding	Condition	ANOVA	$p = .782$	$p = .883$
Study 3				
DV	IV	Test	Normality	Homoscedasticity
Final Distance	Condition	Wilcoxon*	N/A	N/A
Average Distance	$\epsilon_{o/u+a/d}$	Spearman's Correlation	N/A	N/A
$\Delta\epsilon_{o/u+a/d}$	Condition	t-test*	$p = .578$	$p = .848$
$\Delta\epsilon_{o/u}$	Condition	t-test*	$p = .402$	$p = .058$
$\Delta\epsilon_{a/d}$	Condition	t-test*	$p = .193$	$p = .540$
Workload	Condition	t-test*	$p = .814$	$p = .951$
Likeability	Condition	t-test*	$p = .057$	$p = .557$
Intelligence	Condition	t-test*	$p = .697$	$p = .416$
Δ Trust	Condition	Wilcoxon*	N/A	N/A
Δ Team Fluency	Condition	t-test*	$p = .732$	$p = .293$
Δ Understanding	Condition	Wilcoxon*	N/A	N/A

*Test is one-tailed.

APPENDIX C

SAFE META ACTIVE LEARNING FOR DEEP BRAIN STIMULATION

C.0.1 DBS Domain

The input parameter space is one-dimensional and consists of the voltage amplitude (v_a). The output space is a score quantifying memory, referred to as the discrimination score (d_s). The objective of this domain is to maximize discrimination area, which is defined as $d_a := d_s * v_a$. More detail on this domain and the data on which the simulation is based can be found at [18].

C.0.2 Definition of Information Gain

We define information gain, I, at time step t as the percent decrease in the error of the objective as described in Equation C.1. $e^{(t)}$ is the error of the objective at time step t. In the high-dimensional domain, $e^{(t)}$ is defined as the mean squared error of the model, $e^{(t)} = \frac{1}{N}\sum_{i-1}^{N}(x^{(i)} - \hat{x^{(i)}})^2$. x is the ground truth state and $\hat{x}$ is the state predicted by the model. In the DBS domain $e^{(t)}$ is defined as the L_1 norm of the predicted optimal parameter and the ground truth optimal parameter ($e^{(t)} = ||d_a - \hat{d}_a||_1$). During offline training, the ground truth can be obtained from the known model.

$$I^{(t+1)} = \frac{e^{(t)} - e^{(t+1)}}{e^{(t)}} \tag{C.1}$$

C.0.3 Baseline Acquisition Functions

Maximizing Diversity

We compare our method to the acquisition function, maximizing diversity, presented by [33]. Here, $u^{(i)}$ and $x^{(i)}$ are states and actions that the robot has previously experienced, and f is

the current dynamics model.

$$u^* =_{u \in \mathbb{U}} \sum_{i=1}^{N} u - u^{(i)}{}_1 + \beta \hat{T}_\psi \left(x, u\right) - x^{(i)}{}_1 \tag{C.2}$$

Maximizing Uncertainty

This active learning metric described by [32] quantifies the uncertainty in the output of the model for each training example as described in Equation C.3. Here, $\hat{T}_z(u))$ is the z^{th} dynamics model and $\bar{T}$ is the average across models z.

$$u^* =_{u \in \mathbb{U}} \frac{1}{Z} \sum_{z=1}^{Z} \bar{T} - \hat{T}_z(u))_1 \tag{C.3}$$

C.0.4 Additional Details on Mixed-Integer Linear Program Formulation

In this section we provide additional details on the linearization of our probability constraints and objective function for integration into a linear programming formulation.

Dynamics Model Representation

In practice, in the high-dimensional domain we find that $\hat{T}_\psi$ can be represented as a single-layer perceptron (i.e. linear regression) which advantageously is computationally efficient in domains that require fast computation times. We adopt a multi-layer perceptron with ReLU activations in the DBS domain. Our inferred dynamics therefore evolve according Γ. A describes the evolution of a state with no input, B the change in the state due to an action at time t, $\vec{u}^{(t)}$, and I is the identity matrix.

$$\vec{X}^{(t+1:t+T)} = \beta \vec{x}^{(t)} + \Gamma \vec{\mathcal{U}}^{(t:t+T)} \tag{C.4}$$

$$\beta = \begin{bmatrix} A^2 & A^3 & \dots & A^T \end{bmatrix} \tag{C.5}$$

$$\Gamma = \begin{bmatrix} AB & B & 0 & 0 & \dots 0 \\ A^2B & AB & B & 0 & \dots 0 \\ \vdots & \ddots & & & \\ A^{T-1}B & A^{T-2}B & A^{T-3} & \dots & \dots B \end{bmatrix} \tag{C.6}$$

Linearization of Probability Constraints

To include our safety constraints in a mixed-integer linear programming formulation, we remove the non-linearities via conservative assumptions and other techniques. Our safety constraints are defined in Equation C.7 and the dynamics evolve according to equations Equation C.4-Equation C.6. d represents the d^{th} row and j represents the columns.

$$\left\| \Phi^{-1}(1-\epsilon_d) \sqrt{\sum_j \sigma_{d,j}^2 x_j^{(t)^2} + \sum_j \sigma_{d,j}^2 \mathcal{U}_j^{(t:t+T)^2}} + \Gamma \vec{\mathcal{U}}^{(t:t+T)^2} - \Delta_d^{(t:t+T)} \right\|_1 < r_d \tag{C.7}$$

The following conservative assumption is made to linearize the sum of squares in Equation C.7.

$$0 \leq \sqrt{\sum_j \sigma_{d,j}^2 x_j^2} \leq \sum_j \sigma_{d,j} |x_j| \tag{C.8}$$

We utilize the binary decision variable, $\delta_\epsilon \in \{0, 1\}$ as a probability selector variable to linearize the absolute value in Equation C.7. M represents a large positive number and

$\bar{\Gamma}$ is the point estimate of the dynamics. This linearization technique combined with the conservative assumption in Equation C.8 results in the following linear equations (Equation C.9-Equation C.11) which can be integrated into a mixed-integer linear programming formulation. E is the set of "probability levels", e.g., $E = \{0.75, 0.8, ...\}$ where $\min E$ defines the minimum enforced probability of safety.

$$- M\delta_\epsilon - \Phi^{-1}(1-\epsilon_p)\sum_j \sigma_{d,j}\tilde{\mathcal{U}}_j^{(t:t+T)} - \bar{\Gamma}\vec{\mathcal{U}}^{(t:t+T)} < \vec{r_d} + \Delta_d^{(t:t+T)} + \sum_j \sigma_{d,j}|x_j^{(t)}| \tag{C.9}$$

$$- M\delta_\epsilon + \Phi^{-1}(1-\epsilon_p)\sum_j \sigma_{d,j}\tilde{\mathcal{U}}_j^{(t:t+T)} + \bar{\Gamma}\vec{\mathcal{U}}^{(t:t+T)} < r_d - \Delta_d^{(t:t+T)} - \sum_j \sigma_{d,j}|x_j^{(t)}| \tag{C.10}$$

$$\sum_{p\in E} \delta_{\epsilon_{p,d}} = |E| - 1, \forall d \in D \tag{C.11}$$

Variance Estimation

We compute uncertainty of our network via bootstrapping. We follow the method proposed in [241] and randomly redraw bootstrap training samples with replacement from our set of training data. This technique has been verified by [242] to be an effective method for approximating the uncertainty of neural networks. The components of σ are calculated according to Equation C.12. Here $\bar{x}$ is the average of the bootstrapped networks and $x_{d,j}^{(b)}$ a single bootstrapped network. B is the number of bootstrapped networks.

$$\sigma_{d,j} = \sqrt{\frac{\sum_{b=1}^{B}(\bar{x}_{d,j} - x_{d,j}^{(b)})^2}{B-1}} \tag{C.12}$$

Linearization of Q-Function

Our Q-function includes a non-linear relu activation function which is linearized to be included in the mixed-integer linear programming formulation. Equations Equation C.13-Equation C.15 define the equations for a neural network with relu activation. $\xi = \left[\left[\mathcal{U}^{(t:T)}\right], \left[\vec{Z}\right]\right]$ is the input to the Q-function and ${}^{(l)}\omega_{j,i}$ is the connection between neurons j and i between layers l and $l+1$.

$$Q(\vec{\mathcal{U}}^{(t:T)}, \vec{Z}) = \sum_j {}^{(2)}\omega_{j,d}\, {}^{(2)}o_j, \forall d \in D \tag{C.13}$$

$${}^{(2)}o_i = \sum_j {}^{(1)}\omega_{j,i}\, {}^{(1)}o_j \mathbb{1}_{\left({}^{(1)}o_j \geq 0\right)}, \forall i \tag{C.14}$$

$${}^{(1)}o_i = \sum_j {}^{(0)}\omega_{j,i}\xi_j^{(t)} \tag{C.15}$$

This formulation is linearized in Equation C.16-Equation C.18. $k_i \in \{0, 1\}$ is a binary indicator variable and M represents a large positive number.

$$Mk_i - M + \sum_j {}^{(0)}\omega_{j,i}\xi_j^{(t)} \leq 0 \leq Mk_i + \sum_j {}^{(0)}\omega_{j,i}\xi_j^{(t)} \tag{C.16}$$

$$\sum_j {}^{(0)}\omega_{j,i}\xi_j^{(t)} - M \leq {}^{(1)}o_i \leq \sum_j {}^{(0)}\omega_{j,i}\xi_j^{(t)} + Mk_i \tag{C.17}$$

$$M - Mk_i \geq {}^{(1)}o_i \geq 0, \forall i \tag{C.18}$$

Linearization of "resetting" term

In practice, we find that taking a final, "resetting" action at time $t+T$ by adding $z_3 = \vec{x}^{(T)} - \vec{x}_r$ to minimize the distance between the aircraft's state and $\vec{x^r}$ helps to ensure the aircraft does not loiter along the boundary of safe operation until where a random perturbation could result in failure the aircraft. We linearize the resetting term in our objective function, i.e. the difference between our designated safe state, $\vec{x}_r$ and the final

state x_T by maximizing $-(z^+ + z^-)$ subject to the constraint in Equation C.19. z^+ and z^- are both positive continuous variables.

$$\vec{x}_T - \vec{x}_r = z^+ - z^- \tag{C.19}$$

$$z^+, z^- > 0 \tag{C.20}$$

Linearized Objective

The resultant linearized objective is defined in Equation C.21. In the DBS domain λ_3 is set to 0.

$$\vec{\mathcal{U}}^{(t:T)^*} =_{\vec{\mathcal{U}}^{(t:T)} \in \vec{\boldsymbol{\mathcal{U}}}^{(t:T)}} \Big(\lambda_1 \sum_{j}^{H} ({}^{(2)}\omega_{j,d} * {}^{(2)}o_j - \bar{Q}_\theta(\cdot, \vec{Z})) + \lambda_2 \sum_{p \in E} (1 - \delta_{\epsilon_{p,d}}) \epsilon_{p,d} - \lambda_3 (z_d^- + z_d^+) \Big) \tag{C.21}$$

C.1 Safety

The baselines used in the DBS domain did not have built in safety guarantees. Therefore, when comparing against these baselines, we removed the safety constraints in our algorithm to make the comparison fair.

LAL [94], however, is not an inherently safe method of active learning. To fairly compare this baseline to our method in the high-dimensional domain, we simulate the possible actions that can be selected by LAL and discard those that are not safe (i.e., the actions that take the aircraft out of the cylinder of safety). Therefore, LAL can only select an action considered safe by our definition of safety.

C.1.1 Sensitivity Analysis

	Our Approach		Diversity [33]					Uncertainty [32]		
Parameter Setting	5	30	1	2	3	4	5	2	3	4
Average Information Gain (SD)	.196 (.37)	**.39 (.23)**	.29 (.44)	.30 (.30)	.31 (.30)	.26 (.22)	.26 (.22)	.22 (.45)	.25 (.17)	.33 (.17)
Average Computation Time (s) (SD)	.12 (.03)	**.146 (.01)**	.17 (.03)	.18 (.04)	.19 (.04)	.21 (.04)	0.22 (.03)	.16 (.04)	.21 (.05)	.26 (.05)

Table C.1: Average information gain for our approach compared to that by [33] and [32]. We vary the number of previous samples in the diversity maximization problem from one to five and the number of bootstrapped models from two to four in maximizing uncertainty heuristic. We vary the number of hidden neurons in our meta-learned Q-function. We **bold** the setting of our algorithm that outperforms our baselines across *all* hyperparameter settings tested.

To robustly evaluate our method compared to the baselines, we vary the hyperparameters of the approach by [33] for maximizing diversity and our meta-learned function to show that our function is robust and is still superior despite changes in hyperparameters. The results of this hyperparameter sweep are shown in Table C.1. The hyper-parameter we vary for maximizing diversity is the number of previous training samples that we compare to. The information gain increases as the number of samples increases up to a point at which the selected sample tends to converge to the mean of the previously collected samples, causing the information gain to fall. We vary the number of hidden neurons in our Q-function as the hyper-parameter of interest as it governs the trade-off between computational speed and function approximation power.

In our approach, the addition of a hidden neuron adds an additional integer variable, resulting in an increase in computation time as demonstrated in Table C.1. However, an addition of 25 neurons only increase the computation time by 3.7% and provides a 50% increase in information gain. In comparison, [33]'s approach results in increase in information gain 4% while trading off a 6% loss in computation time. Likewise, increasing the number of bootstrapped models in [32]'s approach results in a 35% increase in information gain and

a 37.5% increase in computation time. Therefore, in our approach we are able to gain more information without large increases in computation time.

C.1.2 Additional Results

We present results for each damage condition in the high-dimensional domain comparing our approach to the baseline acquisition functions diversity [33] and epistemic uncertainty [32]. Our approach outperforms the baselines for each damage condition.

C.1.3 Hyperparameters

The hyperparameters employed for each domain are listed in the Table C.2. We show the learning rates for each domain which were determined via experimentation to be effective values. The size of the LSTM hidden layer and each of the two layers of the Q-function is also presented. τ is the soft update coefficient for updating the target network.

	DBS
Learning Rate	1e-4
LSTM Hidden Layer Size	20
Q function Layer 1 Size	64
Q function Layer 2 Size	16
Soft Update (τ)	.001
Exploration noise	Gaussian

Table C.2: Hyperparameters for the DBS and high-dimensional domains.

REFERENCES

[1] M. L. Schrum, E. Hedlund, and M. C. Gombolay, "Improving robot-centric learning from demonstration via personalized embeddings," Oct. 2021.

[2] M. L. Schrum, N. M. Erin Hedlund, and M. C. Gombolay, "Mind meld: Personalized meta-learning for robot-centric imitation learning," *ACM/IEEE International Conference on Human-Robot Interaction*, 2022.

[3] M. L. Schrum, E. Hedlund-Botti, and M. C. Gomoblay, "Reciprocal mind meld: Improving learning from demonstration via personalized, reciprocal teaching," *Conference on Robot Learning*, 2022.

[4] M. L. Schrum, E. Sumner, M. C. Gombolay, and A. Best, *Maveric: A data-driven approach to personalized autonomous driving*, 2023.

[5] M. Schrum, M. J. Connolly, E. Cole, M. Ghetiya, R. Gross, and M. C. Gombolay, "Meta-active learning in probabilistically safe optimization," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 713–10 720, 2022.

[6] M. Nonić and M. Šijačić-Nikolić, "Genetic diversity: Sources, threats, and conservation," in *Life on Land*, W. Leal Filho, A. M. Azul, L. Brandli, A. Lange Salvia, and T. Wall, Eds. Cham: Springer International Publishing, 2021, pp. 421–435, ISBN: 978-3-319-95981-8.

[7] F. A. Lucini, F. Morone, M. S. Tomassone, and H. A. Makse, "Diversity increases the stability of ecosystems," *PLoS ONE*, vol. 15, 4 Apr. 2020.

[8] A. Forsman, *Rethinking phenotypic plasticity and its consequences for individuals, populations and species*, Oct. 2015.

[9] D. Bavelier, D. M. Levi, R. W. Li, Y. Dan, and T. K. Hensch, "Removing brakes on adult brain plasticity: From molecular to behavioral interventions," vol. 30, Nov. 2010, pp. 14 964–14 971.

[10] L. Quintana-Murci, "Genetic and epigenetic variation of human populations: An adaptive tale," *Comptes Rendus Biologies*, vol. 339, no. 7, pp. 278–283, 2016, Trajectories of genetics, 150 years after Mendel / Trajectoire de la génétique, 150 après Mendel Guest Editors / Rédacteurs en chef invités : Bernard Dujon, Georges Pelletier.

[11] C. Clabaugh and M. Matarić, "Robots for the people, by the people: Personalizing human-machine interaction," *Science Robotics*, vol. 3, no. 21, eaat7451, 2018. eprint: https://www.science.org/doi/pdf/10.1126/scirobotics.aat7451.

[12] S. Schneider and F. Kummert, "Comparing robot and human guided personalization: Adaptive exercise robots are perceived as more competent and trustworthy," *International Journal of Social Robotics*, vol. 13, pp. 169–185, 2 Apr. 2021.

[13] A. Kalinowska, A. Prabhakar, K. Fitzsimons, and T. Murphey, "Ergodic imitation: Learning from what to do and what not to do."

[14] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, *Recent Advances in Robot Learning from Demonstration*. 2020, vol. 3, pp. 297–330, ISBN: 1008190632.

[15] L. Chen, S. Jayanthi, R. Paleja, D. Martin, V. Zakharov, and M. Gombolay, "Fast lifelong adaptive inverse reinforcement learning from demonstrations," *Conference on Robot Learning*, 2022.

[16] D. S. Brown, W. Goo, and S. Niekum, "Better-than-demonstrator imitation learning via automatically-ranked demonstrations," Jul. 2019.

[17] M. Cakmak, C. Chao, and A. L. Thomaz, "Designing interactions for robot active learners," *IEEE Transactions on Autonomous Mental Development*, vol. 2, pp. 108–118, 2 May 2010.

[18] O. Ashmaig, M. Connolly, R. E. Gross, and B. Mahmoudi, "Bayesian Optimization of Asynchronous Distributed Microelectrode Theta Stimulation and Spatial Memory," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2018-July, pp. 2683–2686, 2018.

[19] D. Moore, R. Currano, M. Shanks, and D. Sirkin, "Defense against the dark cars: Design principles for griefing of autonomous vehicles," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20, Cambridge, United Kingdom: Association for Computing Machinery, 2020, pp. 201–209, ISBN: 9781450367462.

[20] A. Shin, J. H. Oh, and J. Lee, "Apprentice of oz: Human in the loop system for conversational robot wizard of oz," in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '19, Daegu, Republic of Korea: IEEE Press, 2020, pp. 516–517, ISBN: 9781538685556.

[21] M. J.-Y. Chung, A. Pronobis, M. Cakmak, D. Fox, and R. P. Rao, "Exploring the potential of information gathering robots," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, ser. HRI'15 Extended Abstracts, Portland, Oregon, USA: Association for Computing Machinery, 2015, pp. 29–30, ISBN: 9781450333184.

[22] X. Wang, K. Lee, K. Hakhamaneshi, P. Abbeel, and M. Laskin, "Skill preferences: Learning to extract and execute robotic skills from human feedback," *ArXiv*, vol. abs/2108.05382, 2021.

[23] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[24] R. Dromnelle, B. Girard, E. Renaudo, R. Chatila, and M. Khamassi, "Coping with the variability in humans reward during simulated human-robot interactions through the coordination of multiple learning strategies," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2020, pp. 612–617.

[25] L. Chen, "Robot learning from heterogeneous demonstration a dissertation presented to the academic faculty," 2020.

[26] L. D. Riek, "Wizard of oz studies in hri: A systematic review and new reporting guidelines," *J. Hum.-Robot Interact.*, vol. 1, no. 1, pp. 119–136, 2012.

[27] M. A. Jamal and G.-J. Qi, "Task agnostic meta-learning for few-shot learning."

[28] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," *Conference on Neural Information Processing Systems*, 2018.

[29] X. Sun *et al.*, "Exploring personalised autonomous vehicles to influence user trust," *Cognitive Computation*, vol. 12, no. 6, pp. 1170–1186, Nov. 2020.

[30] F. Ekman, M. Johansson, L.-O. Bligård, M. Karlsson, and H. Strömberg, "Exploring automated vehicle driving styles as a source of trust information," *Transportation Research Part F: Traffic Psychology and Behaviour*, 2019.

[31] A. Andree *et al.*, "Deep brain stimulation electrode modeling in rats," *Experimental Neurology*, vol. 350, p. 113 978, 2022.

[32] T. Hastie, R. Tibshirani, and J. Friedman, "Model Assessment and Selection," in *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, 2017, pp. 249–252.

[33] M. L. Schrum, M. Johnson, M. Ghuy, and M. C. Gombolay, "Four years in review: Statistical practices of likert scales in human-robot interaction studies," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2020, ISBN: 9781450370578.

[34] S. Calinon and A. Billard, "Incremental learning of gestures by imitation in a humanoid robot," in *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2007, pp. 255–262.

[35] R. Paleja and M. Gombolay, "Heterogeneous learning from demonstration," Jan. 2020.

[36] L. Chen, R. Paleja, M. Ghuy, and M. Gombolay, "Joint goal and strategy inference across heterogeneous demonstrators via reward network distillation," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20, Cambridge, United Kingdom: Association for Computing Machinery, 2020, pp. 659–668, ISBN: 9781450367462.

[37] D. Brown, W. Goo, P. Nagarajan, and S. Niekum, "Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 783–792.

[38] M. Laskey *et al.*, "Comparing human-centric and robot-centric sampling for robot deep learning from demonstrations," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 358–365, 2017. arXiv: 1610.00850.

[39] S. Adams, T. Cody, and P. A. Beling, "A survey of inverse reinforcement learning," *Artificial Intelligence Review*, vol. 55, pp. 4307–4346, 6 Aug. 2022.

[40] Y. Gao, H. Xu, J. Lin, F. Yu, S. Levine, and T. Darrell, "Reinforcement learning from imperfect demonstrations," *Conference on Robot Learning*, 2018.

[41] L. Chen, R. R. Paleja, and M. C. Gombolay, "Learning from suboptimal demonstration via self-supervised reward regression," in *Conference on Robot Learning*, 2020.

[42] M. F. Dixon, I. Halperin, and P. Bilokon, "Inverse reinforcement learning and imitation learning," in *Machine Learning in Finance: From Theory to Practice*. Cham: Springer International Publishing, 2020, pp. 419–517, ISBN: 978-3-030-41068-1.

[43] B. Argall, S. Chernova, M. M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, May 2009.

[44] S. Chernova and M. Veloso, "Interactive policy learning through confidence-based autonomy," *Journal of Artificial Intelligence Research*, vol. 34, pp. 1–25, 2009.

[45] S. Ross, G. J. Gordon, and J. A. Bagnell, "No-regret reductions for imitation learning and structured prediction," *Aistats*, vol. 15, pp. 627–635, 2011.

[46] R. Liu, M. C. Gombolay, and S. Balakirsky, "Torwards Unpaired Human-to-Robot Demonstration Translation Learning Novel Tasks," *ICSR Workshop Human Robot Interaction for Space Robotics (HRI-SR)*, 2021.

[47] S. Ross and J. A. Bagnell, "Efficient reductions for imitation learning," *Journal of Machine Learning Research*, vol. 9, pp. 661–668, 2010.

[48] R. Jena, C. Liu, and K. Sycara, "Augmenting GAIL with BC for sample efficient imitation learning," pp. 1–11, 2020.

[49] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, "HG-DAgger: Interactive imitation learning with human experts," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2019-May, pp. 8077–8083, 2019. arXiv: 1810.02890.

[50] M. Laskey *et al.*, "SHIV: Reducing supervisor burden in DAgger using support vectors for efficient learning from demonstrations in high dimensional state spaces," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2016-June, pp. 462–469, 2016.

[51] B. Packard and S. Onta, "A User Study on Learning from Human Demonstration," no. Aiide, pp. 208–214, 2018.

[52] H. Daume and J. Eisner, "Imitation Learning by Coaching," *Conference on Neural Information Processing Systems*, pp. 1–9, 2012.

[53] J. Spencer *et al.*, "Learning from Interventions: Human-robot interaction as both explicit and implicit feedback," 2020.

[54] K. Menda, K. Driggs-Campbell, and M. J. Kochenderfer, "EnsembleDAgger: A Bayesian Approach to Safe Imitation Learning," *IEEE International Conference on Intelligent Robots and Systems*, no. 2, pp. 5041–5048, 2019. arXiv: 1807.08364.

[55] W. Bradley Knox and P. Stone, "TAMER: Training an Agent Manually via Evaluative Reinforcement," in *2008 7th IEEE International Conference on Development and Learning*, 2008, pp. 292–297.

[56] D. H. Grollman, Aude, and G. Billard, "Robot learning from failed demonstrations," *International Journal of Social Robotics*, vol. 4, no. 4, pp. 331–342, 2012.

[57] M. Valko, M. Ghavamzadeh, and A. Lazaric, "Semi-supervised apprenticeship learning," in *Proceedings of the Tenth European Workshop on Reinforcement Learning*,

M. P. Deisenroth, C. Szepesvári, and J. Peters, Eds., ser. Proceedings of Machine Learning Research, vol. 24, Edinburgh, Scotland: PMLR, 2013, pp. 131–142.

[58] B. Burchfiel, C. Tomasi, and R. Parr, "Distance minimization for reward learning from scored trajectories," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016.

[59] A. Sena and M. Howard, "Quantifying teaching behavior in robot learning from demonstration," 2020.

[60] M. Cakmak and L. Takayama, "Teaching people how to teach robots: The effect of instructional materials and dialog design," in *ACM/IEEE International Conference on Human-Robot Interaction*, IEEE Computer Society, 2014, pp. 431–438, ISBN: 9781450326582.

[61] A. Weiss, J. Igelsbock, S. Calinon, A. Billard, and M. Tscheligi, "Teaching a humanoid: A user study on learning by demonstration with hoap-3," in *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 2009, pp. 147–152.

[62] R. Toris, H. B. Suay, and S. Chernova, "A practical comparison of three robot learning from demonstration algorithms," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '12, Boston, Massachusetts, USA: Association for Computing Machinery, 2012, pp. 261–262, ISBN: 9781450310635.

[63] I. Papageorgi, "Positive and negative reinforcement and punishment," in *Encyclopedia of Evolutionary Psychological Science*, T. K. Shackelford and V. A. Weekes-Shackelford, Eds. Cham: Springer International Publishing, 2021, pp. 6079–6081, ISBN: 978-3-319-19650-3.

[64] D. Leyzberg, S. Spaulding, and B. Scassellati, "Personalizing robot tutors to individuals' learning differences," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '14, Bielefeld, Germany: Association for Computing Machinery, 2014, pp. 423–430, ISBN: 9781450326582.

[65] D. Szafir and B. Mutlu, "Pay attention! designing adaptive agents that monitor and improve user engagement," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12, Austin, Texas, USA: Association for Computing Machinery, 2012, pp. 11–20, ISBN: 9781450310154.

[66] G. Gordon *et al.*, "Affective personalization of a social robot tutor for children's second language skills," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16, Phoenix, Arizona: AAAI Press, 2016, pp. 3951–3957.

[67] D. Leyzberg, A. Ramachandran, and B. Scassellati, "The effect of personalization in longer-term robot tutoring," *ACM Transactions on Human-Robot Interaction*, vol. 7, 3 Dec. 2018.

[68] L. Eboli, G. Mazzulla, and G. Pungillo, "How drivers' characteristics can affect driving style," *Transportation Research Procedia*, vol. 27, pp. 945–952, 2017, 20th EURO Working Group on Transportation Meeting, EWGT 2017, 4-6 September 2017, Budapest, Hungary.

[69] O. Taubman - Ben-Ari and V. Skvirsky, "The multidimensional driving style inventory a decade later: Review of the literature and re-evaluation of the scale," *Accident Analysis & Prevention*, vol. 93, pp. 179–188, 2016.

[70] Z. Ma and Y. Zhang, "Investigating the effects of automated driving styles and driver's driving styles on driver trust, acceptance, and take over behaviors," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 64, no. 1, pp. 2001–2005, 2020. eprint: https://doi.org/10.1177/1071181320641484.

[71] C. Basu, Q. Yang, D. Hungerman, M. Sinahal, and A. D. Draqan, "Do you want your autonomous car to drive like you?" In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 417–425.

[72] H. Bellem, T. Schönenberg, J. F. Krems, and M. Schrauf, "Objective metrics of comfort: Developing a driving style for highly automated vehicles," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 41, pp. 45–54, 2016.

[73] J. M. Harris and P. B. Norman, "The aggressive driving behavior scale: Developing a self-report measure of unsafe driving practices," *North American Journal of Psychology*, vol. 5, pp. 193–202, 2003.

[74] N. M. Yusof, J. Karjanto, J. Terken, F. Delbressine, M. Z. Hassan, and M. Rauterberg, "The exploration of autonomous vehicle driving styles: Preferred longitudinal, lateral, and vertical accelerations," in *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Ann Arbor, MI, USA: Association for Computing Machinery, 2016, pp. 245–252, ISBN: 9781450345330.

[75] J. Karlsson, S. van Waveren, C. Pek, I. Torre, I. Leite, and J. Tumova, "Encoding human driving styles in motion planning for autonomous vehicles," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 1050–1056.

[76] J. Iskander *et al.*, "From car sickness to autonomous car sickness: A review," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 62, pp. 716–726, 2019.

[77] J. Reason and J. J. Brand, *Motion sickness / J. T. Reason, J. J. Brand*. Academic Press London ; New York, 1975, xi, 310 p. : ISBN: 0125840500.

[78] B. Keshavarz and H. Hecht, "Validating an efficient method to quantify motion sickness," *Human factors*, vol. 53, no. 4, pp. 415–426, 2011.

[79] M. Kuderer, S. Gulati, and W. Burgard, "Learning driving styles for autonomous vehicles from demonstration," *Internation Conference on Robotics and Automation*, pp. 2641–2646, 2015.

[80] D. Bolduc, L. Guo, and Y. Jia, "Modeling and characterization of driving styles for adaptive cruise control in personalized autonomous vehicles," *Dynamic Systems and Control Conference*, Oct. 2017, V001T44A004. eprint: https://asmedigitalcollection.asme.org/DSCC/proceedings-pdf/DSCC2017/58271/V001T44A004/2375801/v001t44a004-dscc2017-5277.pdf.

[81] Y. Feng and X. Yan, "Support vector machine based lane-changing behavior recognition and lateral trajectory prediction," *Computational Intelligence and Neuroscience*, vol. 2022, p. 3 632 333, May 2022.

[82] J. Ling, J. Li, K. Tei, and S. Honiden, "Towards personalized autonomous driving: An emotion preference style adaptation framework," *IEEE International Conference on Agents*, pp. 47–52, 2021.

[83] R. Feingold-Polak and S. Levy-Tzedek, "Personalized human robot interaction in the unique context of rehabilitation," in *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '21, Utrecht, Netherlands: Association for Computing Machinery, 2021, pp. 126–127, ISBN: 9781450383677.

[84] A. Tapus, C. Ţăpuş, and M. J. Matarić, "User-robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy," *Intelligent Service Robotics*, vol. 1, pp. 169–183, 2 2008.

[85] B. Irfan *et al.*, "Personalised socially assistive robot for cardiac rehabilitation: Critical reflections on long-term interactions in the real world," *User Modeling and User-Adapted Interaction*, 2022.

[86] D. Francois, D. Polani, and K. Dautenhahn, "Towards socially adaptive robots: A novel method for real time recognition of human-robot interaction styles," in *Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots*, 2008, pp. 353–359.

[87] Y. Sui, A. Gotovos, J. W. Burdick, and A. Krause, "Safe exploration for optimization with Gaussian processes," *32nd International Conference on Machine Learning, ICML 2015*, vol. 2, pp. 997–1005, 2015.

[88] M. Turchetta, F. Berkenkamp, and A. Krause, "Safe exploration in finite Markov decision processes with Gaussian processes," *Advances in Neural Information Processing Systems*, no. Nips, pp. 4312–4320, 2016. arXiv: 1606.04753.

[89] C. Zimmer, M. Meister, and D. Nguyen-Tuong, "Safe active learning for time-series modeling with Gaussian processes," *Advances in Neural Information Processing Systems*, vol. 2018-December, no. NeurIPS, pp. 2730–2739, 2018.

[90] R. Burbidge, J. J. Rowland, and R. D. King, "Active Learning for Regression Based on Query by Committee," *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, pp. 209–218, 2007.

[91] M. Hasenjager and H. Ritter, "Active Learning with Local Models 1 Introduction," *Neural Processing Letters*, pp. 107–117, 1998.

[92] W. Cai, M. Zhang, and Y. Zhang, "Batch mode active learning for regression with expected model change," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 7, pp. 1668–1681, 2017.

[93] P. Bachman, A. Sordoni, and A. Trischler, "Learning Algorithms for Active Learning," 2016. arXiv: arXiv:1708.00088v1.

[94] K. Konyushkova, S. Raphael, and P. Fua, "Learning Active Learning from Data," no. Conference on Neural Information Processing Systems (Nips), 2017.

[95] M. Volpp *et al.*, "Meta-learning acquisition functions for transfer learning in bayesian optimization," Apr. 2019.

[96] Y. Geifman and R. El-Yaniv, "Deep Active Learning with a Neural Architecture Search," no. NeurIPS, 2018. arXiv: 1811.07579.

[97] K. Pang, M. Dong, Y. Wu, and T. Hospedales, "Meta-Learning Transferable Active Learning Policies by Deep Reinforcement Learning," pp. 1–8, 2018. arXiv: 1806.04798.

[98] S. Schaal, "Learning from demonstration," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. Jordan, and T. Petsche, Eds., vol. 9, MIT Press, 1997.

[99] S. Chernova and A. L. Thomaz, *Robot Learning from Human Teachers*. Morgan amp; Claypool Publishers, 2014, ISBN: 1627051996.

[100] T. Osa, G. Neumann, and J. Peters, "An Algorithmic Perspective on Imitation Learning," vol. 7, no. 1, pp. 1–179, 2018.

[101] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[102] J. Berggren, "Performance Evaluation of Imitation Learning Algorithms with Human Experts," 2019.

[103] "Automatically constructing control systems by observing human behaviour," *Second International Inductive Logic Programming Workshop*, 1992.

[104] R. Paleja and M. Gombolay, "Inferring personalized bayesian embeddings for learning from heterogeneous demonstration," *arXiv*, 2019. arXiv: 1903.06047.

[105] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," 2017. arXiv: arXiv:1703.03400v3.

[106] E. F. Camacho and C. A. Bordons, *Model Predictive Control in the Process Industry*. Berlin, Heidelberg: Springer-Verlag, 1997, ISBN: 3540199241.

[107] J. M. Snider, "Automatic steering methods for autonomous automobile path tracking," 2009.

[108] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in Neural Information Processing Systems*, pp. 2180–2188, 2016. arXiv: 1606.03657.

[109] M. L. Schrum, E. Hedlund, and M. C. Gombolay, *Improving Robot-Centric Learning from Demonstration via Personalized Embeddings*, _eprint: 2110.03134, 2021.

[110] S. Shah, D. Dey, C. Lovett, and A. Kapoor, *Airsim: High-fidelity visual and physical simulation for autonomous vehicles*, 2017. arXiv: 1705.05065 [cs.RO].

[111] S. Karaman and E. Frazzoli, *Incremental sampling-based algorithms for optimal motion planning*, 2010. arXiv: 1005.0416 [cs.RO].

[112] S. Ross *et al.*, "Learning monocular reactive UAV control in cluttered natural environments," in *2013 IEEE International Conference on Robotics and Automation*, 2013, pp. 1765–1772.

[113] M. L. Schrum, M. Johnson, M. Ghuy, and M. C. Gombolay, "Four years in review: Statistical practices of likert scales in human-robot interaction studies," *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 43–52, 2020.

[114] S. Salvador and P. Chan, "FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space," 2004.

[115] J.-Y. Jian, A. Bisantz, and C. Drury, "Foundations for an Empirically Determined Scale of Trust in Automated Systems," *International Journal of Cognitive Ergonomics*, vol. 4, pp. 53–71, 2000.

[116] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Human mental workload*, vol. 1, no. 3, pp. 139–183, 1988.

[117] C. Bartneck, E. Croft, and D. Kulic, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[118] W. N. Dudley, R. Wickham, and N. Coombs, "An introduction to survival statistics: Kaplan-meier analysis," *Journal of the advanced practitioner in oncology*, vol. 7, no. 1, pp. 91–100, 2016.

[119] G. Troisi, "Humanization builds trust: The effect of human-like chatbots on the willingness to disclose personal information online," *Thesis: Luiss University. Department of Business and Management*,

[120] M. Cakmak and M. Lopes, "Algorithmic and human teaching of sequential decision tasks," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, ser. AAAI'12, Toronto, Ontario, Canada: AAAI Press, 2012, pp. 1536–1542.

[121] A. Alissandrakis, C. L. Nehaniv, and K. Dautenhahn, "Solving the correspondence problem in robotic imitation across embodiments: Synchrony, perception and culture in artifacts," in *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions*, C. L. Nehaniv and K. Dautenhahn, Eds. Cambridge University Press, 2007, pp. 249–274.

[122] G. Hoffman, "Evaluating Fluency in Human-Robot Collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 209–218, 2019.

[123] K. Grammer and E. Oberzaucher, "Our preferences: Why we like what we like," in *Essential Building Blocks of Human Nature*, U. J. Frey, C. Störmer, and K. P. Willführ, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 95–108, ISBN: 978-3-642-13968-0.

[124] H. H. van Huysduynen, J. Terken, J.-B. Martens, and B. Eggen, "Measuring driving styles: A validation of the multidimensional driving style inventory," in *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, New York, NY, USA: Association for Computing Machinery, 2015, pp. 257–264, ISBN: 9781450337366.

[125] M. Hasenjäger and H. Wersing, "Personalization in advanced driver assistance systems and autonomous vehicles: A review," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 1–7.

[126] M. Hoedemaeker, "Driving behaviour with acc and the acceptance by individual drivers," in *ITSC2000. 2000 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.00TH8493)*, 2000, pp. 506–509.

[127] M. Ebnali, R. Lamb, R. Fathi, and K. Hulme, "Virtual reality tour for first-time users of highly automated cars: Comparing the effects of virtual environments with different levels of interaction fidelity," *Applied Ergonomics*, vol. 90, p. 103 226, 2021.

[128] J. K. Choi and Y. G. Ji, "Investigating the importance of trust on adopting an autonomous vehicle," *International Journal of Human–Computer Interaction*, vol. 31, no. 10, pp. 692–702, 2015. eprint: https://doi.org/10.1080/10447318.2015.1070549.

[129] F. M. Poó and R. D. Ledesma, "A study on the relationship between personality and driving styles," *Traffic Injury Prevention*, vol. 14, no. 4, pp. 346–352, 2013, PMID: 23531257. eprint: https://doi.org/10.1080/15389588.2012.717729.

[130] S. Nasiriany, H. Liu, and Y. Zhu, "Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks," *International Conference on Robotics and Automation*, 2021.

[131] I. Bae, J. H. Kim, J. Moon, and S. Kim, "Lane change maneuver based on bezier curve providing comfort experience for autonomous vehicle users," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 2272–2277.

[132] M. Schrum, E. Sumner, M. C. Gombolay, and A. Best, *A data-driven approach to personalized autonomous driving*, https://osf.io/uvrzh/?view_only=f90e7afe0242475abf22dfee6b6233d Sep. 2022.

[133] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[134] A. J. Cooper, L. D. Smillie, and P. J. Corr, "A confirmatory factor analysis of the Mini-IPIP five-factor model personality scale," *Personality and Individual Differences*, vol. 48, no. 5, pp. 688–691, 2010.

[135] J.-Y. Jian, A. Bisantz, and C. Drury, "Foundations for Empirically Determined Scale of Trust in Automated Systems," *International Journal of Cognitive Ergonomics*, vol. 4, pp. 53–71, 1998.

[136] C. Tennant, "Exploring emerging public attitudes towards autonomous vehicles," in *Science Cultures in a Diverse World: Knowing, Sharing, Caring*, B. Schiele, X. Liu, and M. W. Bauer, Eds. Springer Singapore, 2021, pp. 253–266, ISBN: 978-981-16-5379-7.

[137] B. D. Adams, L. E. Bruyn, S. Honde, and P. Angelopoulos, "Trust in automated systems," no. June, p. 136, 2003.

[138] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[139] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The Robotic Social Attributes Scale (RoSAS): Development and Validation," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. Part F1271, pp. 254–262, 2017.

[140] S. E. Lee, E. C. B. Olsen, W. W. Wierwille, and M. Goodman, *A comprehensive examination of naturalistic lane-changes*, 2004.

[141] E. M. Szumska and R. Jurecki, "The effect of aggressive driving on vehicle parameters," *Energies*, vol. 13, no. 24, 2020.

[142] Y. Xu, S. Bao, and A. K. Pradhan, "Modeling drivers' reaction when being tailgated: A random forests method," *Journal of Safety Research*, vol. 78, pp. 28–35, 2021.

[143] Y. Yu, Y. Zhao, D. Li, J. Zhang, and J. Li, "The relationship between big five personality and social well-being of chinese residents: The mediating effect of social support," *Frontiers in Psychology*, vol. 11, Mar. 2021.

[144] G. Hoffman and X. Zhao, "A primer for conducting experiments in human–robot interaction," *J. Hum.-Robot Interact.*, vol. 10, no. 1, Oct. 2020.

[145] I. Asimov, *I, Robot* (Doubleday science fiction). Bantam Books, 1950, ISBN: 9780451018854.

[146] C. V. Dang, M. Jun, Y.-B. Shin, J.-W. Choi, and J.-W. Kim, "Application of modified asimov's laws to the agent of home service robot using state, operator, and result (soar)," *International Journal of Advanced Robotic Systems*, vol. 15, no. 3, p. 1 729 881 418 780 822, 2018. eprint: https://doi.org/10.1177/1729881418780822.

[147] S. Posporelis, A. S. David, K. Ashkan, and P. Shotbolt, "Deep brain stimulation of the memory circuit: Improving cognition in alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 64, no. 2, pp. 337–347, 2018.

[148] S. Yan, K. Chaudhuri, and T. Javidi, "Active learning from imperfect labelers," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016.

[149] C. Zhang and K. Chaudhuri, "Active learning from weak and strong labelers," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28, Curran Associates, Inc., 2015.

[150] P. Ren *et al.*, "A survey of deep active learning," *arXiv*, 2020. arXiv: 2009.00236.

[151] L. Wang, E. A. Theodorou, and M. Egerstedt, "Safe Learning of Quadrotor Dynamics Using Barrier Certificates," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2460–2465, 2018.

[152] K. Nguyen, "Converting pompds into mdps using history representation," *Technical Report.*, 2021.

[153] L. Blackmore, M. Ono, and B. C. Williams, "Chance-constrained optimal path planning with obstacles," *IEEE Transactions on Robotics*, vol. 27, pp. 1080–1094, 6 Dec. 2011.

[154] I. Griva, S. G. Nash, and A. Sofer. Society for Industrial Mathematics, 2009.

[155] E. Timmons *et al.*, "Information-driven and risk-bounded autonomy for scientist avatars."

[156] S. Dai, S. Schaffert, A. Jasour, A. Hofmann, and B. Williams, "Chance constrained motion planning for high-dimensional robots," in *2019 International Conference on Robotics and Automation (ICRA)*, IEEE, 2019, pp. 8805–8811.

[157] M. Ono and B. C. Williams, "Iterative risk allocation: A new approach to robust model predictive control with a joint chance constraint," in *2008 47th IEEE Conference on Decision and Control*, IEEE, 2008, pp. 3427–3432.

[158] M. Ono, B. C. Williams, and L. Blackmore, "Probabilistic planning for continuous dynamic systems under bounded risk," *Journal of Artificial Intelligence Research*, vol. 46, pp. 511–577, 2013.

[159] M. Ono, M. Pavone, Y. Kuwata, and J. Balaram, "Chance-constrained dynamic programming with application to risk-aware robotic space exploration," *Autonomous Robots*, vol. 39, no. 4, pp. 555–571, 2015.

[160] H. Zhu and J. Alonso-Mora, "Chance-constrained collision avoidance for mavs in dynamic environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 776–783, 2019.

[161] M. Ganger, E. Duryea, and W. Hu, "Double sarsa and double expected sarsa with shallow and deep learning," *Journal of Data Analysis and Information Processing*, vol. 4, no. 4, pp. 159–176, 2016.

[162] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016.

[163] M. L. Schrum, M. Ghuy, E. Hedlund-Botti, M. Natarajan, M. J. Johnson, and M. C. Gombolay, "Concerning trends in likert scale usage in human-robot interaction: Towards improving best practices," *J. Hum.-Robot Interact.*, Nov. 2022, Just Accepted.

[164] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 22 140, pp. 55–55, 1932.

[165] A. Joshi, S. Kale, S. Chandel, and D. Pal, "Likert Scale: Explored and Explained," *British Journal of Applied Science & Technology*, vol. 7, no. 4, pp. 396–403, 2015.

[166] R. Mittu, D. Sofge, A. Wagner, and W. F. Lawless, *Robust intelligence and trust in autonomous systems*. 2016, pp. 1–270, ISBN: 9781489976680.

[167] J. Carifio and R. Perla, "Resolving the 50-year debate around using and misusing Likert scales," *Med Educ*, vol. 42, no. 12, pp. 1150–1152, 2008.

[168] J. Carifio and R. J. Perla, "Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes," *Journal of Social Sciences*, vol. 3, no. 3, pp. 106–116, 2007.

[169] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory*, 3rd. New York, New York, USA: McGraw-Hil, 1994.

[170] J. A. Gliem and R. R. Gliem, "Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales," in *Midwest Research to Practice Conference in Adult, Continuing, and Community Education*, Columbus, 2003, pp. 82–88.

[171] T. Verhagen, B. van den Hooff, and S. Meents, "Toward a better use of the semantic differential in IS research: An integrative framework of suggested action," *Journal of the Association of Information Systems*, vol. 16, no. 2, pp. 108–143, 2015.

[172] O. Friborg, M. Martinussen, and J. H. Rosenvinge, "Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience," *Personality and Individual Differences*, vol. 40, no. 5, pp. 873–884, 2006.

[173] S. Hart and L. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Human Mental Workload*, vol. 43, no. 5, pp. 138–179, 1988.

[174] G. B. Reid, S. S. Potter, and J. R. Bressler, "Subjective Workload Assessment Technique (SWAT): A User's Guide," *ARMSTRONG AEROSPACE MEDICAL RESEARCH LABORATORY*, p. 115, 1989.

[175] M. S. Matell and J. Jacoby, "Is there an optimal number of alternatives for likert scale items? study 1: Reliability and validity," *Educational and Psychological Measurement*, vol. 31, no. 3, pp. 657–674, 1971.

[176] C. C. Preston and A. M. Colman, "Optimal number of response categories in rating scales : reliability , validity , discriminating power , and respondent preferences," *Acta Psychologica*, vol. 104, pp. 1–15, 2000.

[177] D. R. Johnson and J. C. Creech, "Ordinal measures in multiple indicator models: A simulation study of categorization error.," *American Sociological Review*, vol. 48, p. 398, 1983.

[178] H. Wu and S.-o. Leung, "Can Likert Scales be Treated as Interval Scales?— A Simulation Study," *Journal of Social Service Research*, vol. 43, no. 4, pp. 527–532, 2017.

[179] A. W. Bendig, "The Reliability of Self-Ratings as a Function of the Amount of Verbal Anchoring and of the Number of Categories on the Scale," *Journal of Applied Psychology*, vol. 37, no. 1, pp. 38–41, 1953.

[180] J. Lee and I. Paek, "In Search of the Optimal Number of Response Categories in a Rating Scale," *Journal of Psychoeducational Assessment*, vol. 32, no. 7, pp. 663–673, 2014.

[181] L. J. Simms, K. Zelazny, T. F. Williams, and L. Bernstein, "Does the Number of Response Options Matter ? Psychometric Perspectives Using Personality Questionnaire Data," *Psychological Assessment*, vol. 31, no. 4, pp. 557–566, 2019.

[182] A. DeCastellarnau, "A classification of response scale characteristics that affect data quality: a literature review," *Quality and Quantity*, vol. 52, no. 4, pp. 1523–1559, 2018.

[183] S. Y. Y. Chyung, K. Roberts, I. Swanson, and A. Hankinson, "Evidence-based survey design: The use of a midpoint on the likert scale," *Performance Improvement*, vol. 56, pp. 15–23, 10 Nov. 2017.

[184] L. Hall, C. Hume, and S. Tazzyman, "Five Degrees of happiness: Effective Smiley Face Likert scales for evaluating with children," *Proceedings of IDC 2016 - The 15th International Conference on Interaction Design and Children*, pp. 311–321, 2016.

[185] B. Courtenay and C. Weidemann, "The Effects of a "Don't Know" Response on Palmore's Facts on Aging Quizzes," *The Gerontologist*, vol. 2, no. 2, pp. 117–181, 1985.

[186] T. M. Madden and F. J. Klopfer, "The "Cannot Decide" Option in Thurstone-Type Attitude Scales," *Educational and Psychological Measurement*, vol. 38, no. 2, pp. 259–264, 1978.

[187] R. F. Guy and M. Norvell, "The Neutral Point on a Likert Scale," *The Journal of Psychology*, vol. 95, no. 2, pp. 199–204, 1997.

[188] J. A. Krosnick, S. Narayan, and W. R. Smith, "Satisficing in surveys: Initial evidence," *New Directions for Evaluation*, vol. 1996, no. 70, pp. 29–44, 1996.

[189] R. Johns, "One Size Doesn't Fit All: Selecting Response Scales For Attitude Items," *Journal of Elections, Public Opinion and Parties*, vol. 15, no. 2, pp. 237–264, 2005.

[190] B. Weijters, E. Cabooter, and N. Schillewaert, "The effect of rating scale format on response styles: The number of response categories and response category labels," *International Journal of Research in Marketing*, vol. 27, no. 3, pp. 236–247, 2010.

[191] D. F. Alwin and J. A. Krosnick, "The reliability of survey attitude measurement: The influence of question and respondent attributes," *Sociological Methods & Research*, vol. 20, no. 1, pp. 139–181, 1991. eprint: https://doi.org/10.1177/0049124191020001005.

[192] J. A. Krosnick, "Response strategies for coping with the cognitive demands of attitude measures in surveys," *Applied Cognitive Psychology*, vol. 5, no. 3, pp. 213–236, 1991. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350050305.

[193] S. O. Leung and M. L. Xu, "Single-Item Measures for Subjective Academic Performance, Self-Esteem, and Socioeconomic Status," *Journal of Social Service Research*, vol. 39, pp. 511–520, Jul. 2013.

[194] J. R. Rossiter, "The C-OAR-SE procedure for scale development in marketing," *International Journal of Research in Marketing*, vol. 19, pp. 305–335, 2002.

[195] L. Bergkvist and J. R. Rossiter, "The predictive validity of multiple-item versus single-item measures of the same constructs," *Journal of Marketing Research*, vol. 44, no. 2, pp. 175–184, 2007.

[196] A. Diamantopoulos, M. Sarstedt, C. Fuchs, P. Wilczynski, and S. Kaiser, "Guidelines for choosing between multi-item and single-item scales for construct measurement : a predictive validity perspective," *Journal of the Academy of Marketing Science*, vol. 40, no. 3, pp. 434–449, 2012.

[197] A. de Boer *et al.*, "Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life?" *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation*, vol. 13, pp. 311–20, Apr. 2004.

[198] T. Yan and R. Tourangeau, "Fast Times and Easy Questions : The Effects of Age , Experience and Question Complexity on Web Survey Response Times," *Applied Cognitive Psychology*, vol. 68, no. February 2007, pp. 51–68, 2008.

[199] P. Moule, *Making Sense of Research in Nursing, Health and Social Care*. SAGE Publications Ltd, 2015.

[200] H. Schuman and S. Presser, *Questions and Answers in Attitude Surveys*. New York, New York, USA: Academic Press, 1981.

[201] J. Yamaguchi, "Positive versus Negative Wording," *Rasch Measurement Transactions*, vol. 11, 1997.

[202] L. C. Quilty, J. M. Oakman, E. Risko, L. C. Quilty, J. M. Oakman, and E. Risko, "Correlates of the Rosenberg Self-Esteem Scale Method Effects," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 5511, pp. 99–117, 2009.

[203] E. van Sonderen, R. Sanderman, and J. C. Coyne, "Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain," *PLoS ONE*, vol. 8, no. 7, pp. 1–7, 2013.

[204] P. M. Horan, C. Distefano, and R. W. Motl, "Wording Effects in Self-Esteem Scales: Methodological Artifact or Response Style?" *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 10, no. 3, pp. 435–455, 2003.

[205] J. Dawes, "Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales," *International Journal of Market Research*, vol. 50, no. 1, pp. 61–77, 2008.

[206] T. Nemoto and D. Beglar, "Developing Likert-Scale Questionnaires," *JALT2013 Conference Proceedings*, 2013.

[207] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, and S. L. Young, "Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer," *Frontiers in Public Health*, vol. 6, no. June, pp. 1–18, 2018.

[208] T. Kline, "Psychological testing: A practical approach to design and evaluation," in Thousand Oaks, California: SAGE Publications, Inc., 2014, ch. Classical Test Theory: Assumptions, Equations, Limitations, and Item Analyses, pp. 91–106.

[209] R. Hambleton and H. Swaminathan, *Item Response Theory: Principles and Applications*. Springer Science & Business Media, 2013.

[210] N. Wongpakaran and T. Wongpakaran, "Reliability Analysis : Its Application in Clinical Practice," *Chiang Mai University, Thailand*, 2013.

[211] R. E. Yagoda and D. J. Gillan, "You Want Me to Trust a ROBOT? The Development of a Human-Robot Interaction Trust Scale," *International Journal of Social Robotics*, 2012.

[212] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The Robotic Social Attributes Scale (RoSAS): Development and Validation," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. Part F1271, pp. 254–262, 2017.

[213] K. S. Taber, "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," *Research in Science Education*, vol. 48, no. 6, pp. 1273–1296, 2018.

[214] A. M. Gadermann, M. Guhn, B. D. Zumbo, and B. Columbia, "Estimating ordinal reliability for Likert-type and ordinal item response data : A conceptual , empirical , and practical guide," *Practical Assessment, Research & Evaluation*, vol. 17, no. 3, pp. 1–13, 2012.

[215] C. Goforth, *Using and Interpreting Cronbach's Alpha*, 2016.

[216] M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha," *International journal of medical education*, vol. 2, pp. 53–55, 2011.

[217] T. Raykov and G. A. Marcoulides, *Introduction to psychometric theory*. Routledge, 2011.

[218] L. Deng and W. Chan, "Testing the difference between reliability coefficients alpha and omega," *Educational and Psychological Measurement*, vol. 77, pp. 185–203, 2 Apr. 2017.

[219] E. B. Ravinder and A. B. Saraswathi, "Literature review of cronbachalphacoefficient () and mcdonald's omega coefficient ()."

[220] R. A. Asún, K. Rdz-Navarro, and J. M. Alvarado, "Developing Multidimensional Likert Scales Using Item Factor Analysis: The Case of Four-point Items," *Sociological Methods and Research*, vol. 45, no. 1, pp. 109–133, 2016.

[221] P. Samuels, "Advice on Exploratory Factor Analysis," *Centre for Academic Success, Birmingham City University*, no. June, p. 2, 2016.

[222] W. P. Handwerker, "Constructing Likert Scales: Testing the Validity and Reliability of Single Measures of Multidimensional Variables," *Cultural Anthropology Methods*, vol. 8, no. 1, pp. 1–7, 1996.

[223] B. Subramanian, "Likert technique of attitude scale construction in nursing research," *Asian J. Nursing Edu. and Research*, vol. 2, pp. 65–69, Jun. 2012.

[224] B. Lantz, "Equidistance of Likert-Type Scales and Validation of Inferential Methods Using Experiments and Simulations," *Electronic Journal of Business Research Methods*, vol. 11, pp. 16–28, 2013.

[225] D. L. Clason and T. J. Dormody, "Analyzing Data Measured By Individual Likert-Type Items," *Journal of Agricultural Education*, vol. 35, no. 4, pp. 31–35, 1994.

[226] P. A. Bishop and R. L. Herron, "Use and Misuse of the Likert Item Responses and Other Ordinal Measures.," *International journal of exercise science*, vol. 8, no. 3, pp. 297–302, 2015.

[227] I. E. Allen and C. A. Seaman, *Likert Scales and Data Analyses*, 2007.

[228] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance," 1972.

[229] G. E. Meek, C. Ozgur, and K. Dunning, "Comparison of the t vs. Wilcoxon Signed-Rank test for likert scale data and small samples," *Journal of Modern Applied Statistical Methods*, vol. 6, no. 1, pp. 91–106, 2007.

[230] M. J. Nanna, "Analysis of Likert Scale Data in Disability and Medical Rehabilitation Research," *Psychological Methods*, vol. 3, no. 1, pp. 55–67, 1998.

[231] A. J. Vickers, "Comparison of an Ordinal and a Continuous Outcome Measure of Muscle Soreness," *Int J Technol Assess Health Care*, vol. 4, no. 1999, pp. 709–716, 2019.

[232] S. Lee and D. K. Lee, "What is the proper way to apply the multiple comparison test?" *Korean Journal of Anesthesiology*, vol. 71, pp. 353–360, 5 Oct. 2018.

[233] P. C. Austin, M. M. Mamdani, D. N. Juurlink, and J. E. Hux, "Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health," *Journal of Clinical Epidemiology*, vol. 59, no. 9, pp. 964–969, 2006.

[234] H.-Y. Kim, " Statistical notes for clinical researchers: post-hoc multiple comparisons," *Restorative Dentistry & Endodontics*, vol. 40, no. 2, p. 172, 2015.

[235] S. Nakagawa, "A farewell to Bonferroni: The problems of low statistical power and publication bias," *Behavioral Ecology*, vol. 15, no. 6, pp. 1044–1045, 2004.

[236] R. Warner, *Applied Statistics From Bivariate Through Multivariate Techniquese*. Sage Publications, 2012, pp. 1–40.

[237] F. Chiarotti, "Detecting assumption violations in mixed-model analysis of variance," *Ann Ist Super Sanità*, vol. 40, no. 2, pp. 165–171, 2004.

[238] C. R. Blair, "A Reaction to "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analysis of Variance and Covariance"," *Review of Educational Research*, vol. 51, no. 4, pp. 499–507, 1981.

[239] D. Zeevi *et al.*, "Personalized nutrition by prediction of glycemic responses," *Cell*, vol. 163, pp. 1079–1094, 5 Nov. 2015.

[240] N. L. Robinson, T.-N. Hicks, G. Suddrey, and D. J. Kavanagh, *The Robot Self-Efficacy Scale: Robot Self-Efficacy, Likability and Willingness to Interact Increases After a Robot-Delivered Tutorial; The Robot Self-Efficacy Scale: Robot Self-Efficacy, Likability and Willingness to Interact Increases After a Robot-Delivered Tutorial*. 2020, ISBN: 9781728160757.

[241] B. Efron, "Bootstrap Methods: Another Look at the Jackknife Author(s): B. Efron Source : The Annals of Statistics , Vol . 7 , No . 1 (Jan ., 1979), pp . 1-26 Published

by : Institute of Mathematical Statistics Stable URL : https://www.jstor.org/stable/2958830," vol. 7, no. 1, pp. 1–26, 2020.

[242] J. Franke and M. A. Nuemann, "Bootstrapping Neural Networks," *Neural Computation*, vol. 12, no. 8, pp. 1929–1949, 1998.