Mathematical Models of Adaptation in Human-Robot Collaboration

Stefanos Nikolaidis

December 7, 2017

The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Siddhartha Srinivasa, CMU RI
Emma Brunskill, CMU RI
Jodi Forlizzi, CMU HCII
Ariel Procaccia, CMU CSD
David Hsu, NUS

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

CMU-RI-TR-17-71

Abstract

While much work in human-robot interaction has focused on leader-follower teamwork models, the recent advancement of robotic systems that have access to vast amounts of information suggests the need for robots that take into account the quality of the human decision making and actively guide people towards better ways of doing their task. This thesis proposes an equal-partners model, where human and robot engage in a dance of inference and action, and focuses on one particular instance of this dance: the robot adapts its own actions via estimating the probability of the human adapting to the robot. We start with a bounded-memory model of human adaptation parameterized by the human adaptability - the probability of the human switching towards a strategy newly demonstrated by the robot. We then examine more subtle forms of adaptation, where the human teammate adapts to the robot, without replicating the robot's policy. We model the interaction as a repeated game, and present an optimal policy computation algorithm that has complexity linear to the number of robot actions. Integrating these models into robot action selection allows for human-robot mutual-adaptation. Human subject experiments in a variety of collaboration and shared-autonomy settings show that mutual adaptation significantly improves human-robot team performance, compared to one-way robot adaptation to the human.

Acknowledgements

I cannot thank enough my advisor, Sidd Srinivasa. Sidd has been a role-model for me as an academic, a mentor and a person. Thank you for teaching me the importance of mathematical rigor, for your unlimited supply of amazing ideas, and for your constant commitment towards my success. Thank you for pushing me to do this extra 20% for each problem we worked on; the extra mile needed to really solve the problem. Working with you has been a – hopefully more than – once in a lifetime experience and an absolute privilege.

I am grateful to my former advisor at MIT, Julie Shah. Julie introduced me to the wonderful world of algorithmic human-robot collaboration, and my work in the Interactive Robotics Group built the foundations for this thesis. Thank you, Julie, also for showing me the importance of facing every challenge with an enthusiastic, positive attitude.

I would like to express my gratitude to David Hsu, whose constant support and contribution has been a determining factor in this thesis. I am truly inspired by your intellect, technical rigor and dedication in our countless hours of discussion during your visit at CMU, through Skype, emails, and during my visit to Singapore. Thank you, David, also for introducing me to your fantastic M^2AP group.

I am grateful to Ariel Procaccia for sharing your brilliance and for your wicked foresight on what can be solved in polynomial time. Thank you Ariel for helping me accomplish my longtime goal of writing an algorithmic game theory paper. I would also like to thank Jodi Forlizzi for your excellent feedback in designing and running user studies, for putting my thoughts in context and for your unlimited knowledge of related work. Thank you Jodi for always making time to meet with me! I am also truly thankful to Emma Brunskill for your excellent and rigorous feedback, and for emphasizing the importance of robustness in the models.

I have been very fortunate to have some amazing collaborators: Thank you Guy Hoffman for your insightful remarks about the big picture of this work; Mike Koval for knowing the answer to every question I asked, for the late-night discussions about the nitty-gritty details of POMDP solvers, for introducing me to the ADA code and inspiring me with your intellect and character; Swaprava Nath for joining me in the algorithic game-theory journey; Min Chen and Harold Soh for working together on trust modeling and for the great time in Singapore; Anca Dragan for your insights on presenting research ideas; Shervin Javdani and Henny Admoni for our discussions about math, figures and studies; Rachel Holladay, Gilwoo Lee and Oren Salzman for your tough love on my presentations; Clint Liddick for making everything work and for being so much fun to have around; Laura Herlant for your help with ADA and for participating in our New Zealand and Canada adventures; Katharina Muelling for being an awesome co-instructor in "Manipulation Algorithms"; Pyry Matikainen for your web / VR wizardry; Matt Klingensmith and Jennifer King for sharing your expertise; Keyla Cook, Jean Harpley and Suzanne Lyons Muth for being so supportive with all the administrative issues.

I am also grateful to the students that I have been so lucky to mentor and have their support: Billy Zhu for your help in deploying user studies and building ADA infrastructure; Minae Kwon for working together on verbal communication and for your insights on our studies; Anton Kuznetsov for building the survey framework for online studies that has (almost) never failed me through the whole duraction of my graduate studies; Rosario Scalise for working together on weight inference and for being such a great friend; Shen Li for our discussions and the "The Wire" T-shirt!

Being a member of the Personal Robotics Lab has been a fantastic experience. I have been really fortunate to work at such an envigorating and supportive environment. Thank you Aaron J, Aaron W, Aditya, Anca, Ariana, Brian O, Brian H, Chris, Clint, Daqing, Gilwoo, Henny, Jenn, Jimmy, JS, Laura, Mike D, Mike K, Oren, Pras, Pyry, Rachel, Rosario, Shen, Shervin, Shushman and Vinitha! I will miss the lab lunches, weekly meetings, herbathlons / adathlons and karaoke nights!

I would also like to acknowledge the Onassis Foundation as a sponsor and thank them for their support.

Last but not least, I cannot thank enough my parents, Zachos and Efi, and my sister Evelina, for their unconditioned love and support.

Contents

List of Figures

List of Tables

1
Introduction

In collaboration, the success of the team often depends on the ability of team members to coordinate their actions, by reasoning over the beliefs and actions of their teammates. We want to enable robot teammates with this very capability in human-robot teams, e.g., service robots interacting with users at home, manufacturing robots sharing the same physical scape with human mechanics and autonomous cars interacting with drivers and pedestrians.

When it comes to robots operating in isolation, there has been tremendous progress in enabling them to act autonomously by reasoning over the physical state of the world. A manipulator picking up a glass needs to know the position and orientation of the glass on the table, the location of other objects that it should avoid, and the way these objects will move if pushed to the side. More importantly, it needs to reason over the uncertainty in its model of the world and *adapt* its own actions to account for this uncertainty, for instance by looking at the table with its camera, or by moving slowly until it senses the glass in its gripper.

However, humans are not just obstacles that the robot should avoid. They are intelligent agents with their own internal state, i.e., their own goals and expectations about the world and the robot. Their state can change, as they *adapt* themselves to the robot and its actions (Fig. 1.1). Much like in manipulation, a robot interacting with people needs to use this information when choosing its own actions. This requires not only an understanding of human behavior when interacting with robotic systems, but also of the computational challenges and opportunities that arise by enabling this reasoning into deployed systems in the real world.

To address these challenges, we have used insights from behavioral economics to propose scalable models of human behavior and machine learning algorithms to automatically learn these models from data. Integrating these models into probabilistic planning and game-theoretic algorithms has allowed generation of robot actions in

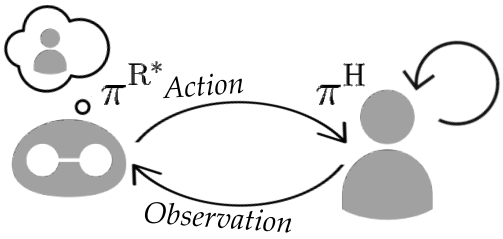


Figure 1.1: The robot maximizes the performance of the human-robot team by executing the optimal policy π^{R^*}. The robot takes *information-seeking* actions that allow estimation of the human policy π^H, but also *communicative* actions that guide π^H towards better ways of doing the task. These actions emerge out of optimizing for π^{R^*}.

a computationally tractable manner.

This thesis has been inspired by recent technical advances in human-robot interaction [Thomaz et al., 2016], and to a large extent it has been made possible by breakthroughs in computational representations of uncertainty [Ong et al., 2010], and in algorithms that have leveraged these representations [Kurniawati et al., 2008]. It has also been inspired by the remarkable results of game-theoretic algorithms in deployed applications [Pita et al., 2009].

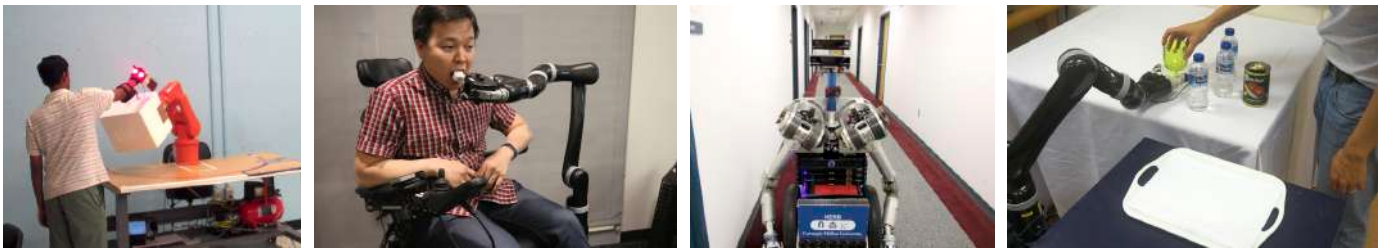
We start by formulating our overarching goal of computing the robot actions that maximize team performance as an optimization problem in a two-player game with incomplete information (chapter 3). In hindsight, our approaches in this thesis reflect the different assumptions and approximations that we made within the scope of this general formulation.

Previous work has assumed a *leader-follower* teamwork model, where the goal of the robot is to follow the human preference (chapter 4). We show that this model is an instance of our general framework by representing the human preference as a reward function, shared by both agents and unknown to the robot.

This thesis then focuses on the case when the robot can indirectly affect human actions as an *equal partner*, by treating the interaction as an underactuated dynamical system (chapter 5). We present a *bounded-memory model* [Nikolaidis et al., 2016, 2017c,a] and a *best-response model* of human behavior [Nikolaidis et al., 2017b], and show that this results in *human-adaptation* to the robot.

Closing the loop between the two results in *mutual adaptation* (chapter 6): The robot builds online a model of the human adaptation by taking information seeking actions, and adapts its own actions in return [Nikolaidis et al., 2016, 2017c,a]. We formalize human-robot mutual adaptation for the collaboration domain, where both human and robot affect the physical state of the world, and for the shared-autonomy domain, where the human simply provides inputs to the robot through a joystick interface. In chapter 7, we generalize the formalism, so that it includes verbal communication from the robot to the human [Nikolaidis et al., 2018].

Each chapter articulates the different assumptions and explains



how these lead to the robot behaviors that we observed in real-time interactions with actual human subjects, in a variety of manufacturing, home environments and assistive care settings (Fig. 1.2).

Figure 1.2: We have applied our research to human-robot collaboration across different robot morphologies and settings: in manufacturing, assistive care, social navigation and at home.

2
Related Work

This thesis builds upon prior work on algorithms for robot adaptation to the human (section 2.1) and human adaptation to the robot (section 2.2), and proposes a human-robot mutual adaptation formalism. We additionally draw upon insights from previous work on verbal communication for the human-human and human-robot teams (section 2.3), and generalize our formalism, so that it incorporates verbal communication from the robot to the human, as well.

2.1 Robot Adaptation

There has been extensive work on one-way robot adaptation to the human. Approaches involve a human expert providing demonstrations to teach the robot a skill or a specific task [Argall et al., 2009, Atkeson and Schaal, 1997, Abbeel and Ng, 2004, Nicolescu and Mataric, 2003, Chernova and Veloso, 2008, Akgun et al., 2012]. Robots have also been able to infer the human preference online through interaction. In particular, partially observable Markov decision process (POMDP) models have allowed reasoning over the uncertainty on the human intention [Doshi and Roy, 2007, Lemon and Pietquin, 2012, Broz et al., 2011]. The MOMDP formulation [Ong et al., 2010] has been shown to achieve significant computational efficiency and has been used in motion planning applications [Bandyopadhyay et al., 2013]. Recent work has also inferred human intention through decomposition of a game task into subtasks for game AI applications. One such study [Nguyen et al., 2011] focused on inferring the intentions of a human player, allowing a non-player character (NPC) to assist the human. Alternatively, Macindoe et al. proposed the partially observable Monte-Carlo cooperative planning system, in which human intention is inferred for a turn-based game [Macindoe et al., 2012]. Nikolaidis et al. proposed a formalism to learn human types from joint-action demonstrations, infer online the type of a new user and compute a robot policy aligned to their preference [Nikolaidis

et al., 2015b]. Simultaneous intent inference and robot adaptation has also been achieved through propagation of state and temporal constraints [Karpas et al., 2015]. Another approach has been the human-robot cross-training algorithm, where the human demonstrates their preference by switching roles with the robot, shaping the robot reward function [Nikolaidis and Shah, 2013]. Although it is possible that the human changes strategies during the training, the algorithm does not use a model of human adaptation that can enable the robot to actively influence the actions of its human partner.

2.2 *Human Adaptation*

There have also been studies in human adaptation to the robot. Previous work has focused on operator training for military, space and search-and-rescue applications, with the goal of reducing the operator workload and operational risk [Goodrich and Schultz, 2007]. Additionally, researchers have studied the effects of repeated interactions with a humanoid robot on the interaction skills of children with autism [Robins et al., 2004], on language skills of elementary school students [Kanda et al., 2004], as well as on users' spatial behavior [Green and Hüttenrauch, 2006]. Human adaptation has also been observed in an assistive walking task, where the robot uses human feedback to improve its behavior, which in turn influences the physical support provided by the human [Ikemoto et al., 2012]. While the changes in the human behavior are an essential part of the learning process, the system does not explicitly reason over the human adaptation throughout the interaction. On the other hand, Dragan and Srinivasa proposed a probabilistic model of the inference made by a human observer over the robot goals, and introduced a motion generating algorithm to maximize this inference towards a predefined goal [Dragan and Srinivasa, 2013a].

Our proposed formalism of human-robot mutual adaptation [1] is an attempt to close the loop between the two lines of research. The robot leverages a human adaptation model parameterized by human adaptability. It reasons probabilistically over the different ways that the human may change the strategy and adapts its own actions to guide the human towards a more effective strategy when possible.

[1] Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddhartha Srinivasa. Formalizing human-robot mutual adaptation: A bounded memory model. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016; Stefanos Nikolaidis, Yu Xiang Zhu, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in shared autonomy. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017c; and Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research (IJRR)*, 2017a

2.3 *Verbal Communication*

In previous work, verbal communication has been frequently used as a mediator of the adaptation process to facilitate communication and resolve conflict. We use insights from studies in verbal communication in human-human teams and human-robot teams, to integrate

verbal communication in the mutual adaptation formalism.

2.3.1 Human-Human Teams

Verbal discourse is a joint activity [Clark, 1994], where participants need to establish a shared understanding of their mutual knowledge base. This shared understanding, also called common ground, can be organized into two types: a communal common group, which represents universal shared knowledge, and personal common groups which represent mutual knowledge gathered from personal experience [Clark, 1994, 1996]. People develop personal common ground by contributing new information, which enables participants in the conversation to reach a mutual belief. This belief, known as grounding [Clark and Schaefer, 1989], indicates that they have understood the information as the speaker intended. Grice [1975] has shown that grounding is achieved when people avoid expending unnecessary effort to convey information.

Previous work has shown that establishing grounding through verbal communication can improve performance, even when combined with other types of feedback. Wang et al. [2013] show that the efficiency of haptic communication was improved only after dyads were first given a learning period in which they could familiarize themselves with the task using verbal communication. Parikh et al. [2014] find that for a more complicated task, verbal feedback coupled with haptic feedback has a significant positive effect on team performance, as opposed to haptic feedback alone. In general, verbalization is more flexible than haptic feedback, since it allows for the communication of more abstract and complex ideas [Eccles and Tenenbaum, 2004], while it can facilitate a shared understanding of the task [Bowers et al., 1998].

However, verbal communication is costly in terms of time and cognitive resources [Eccles and Tenenbaum, 2004]. For example, according to Clark and Brennan [1991], it costs time and effort to formulate coherent utterances, especially when talking about unfamiliar objects or ideas. Receivers also experience costs in receiving and understanding a message; listening and understanding utterances can be especially costly when contextual cues are missing and the listener needs to infer the meaning. Thus, after teams have a shared understanding of the task, it may be beneficial to switch to a less costly mode of communication, such as haptic feedback. In fact, Kucukyilmaz et al. [2013] show that haptic feedback increases a perceived sense of presence and collaboration, making interaction easier. Haptic communication has been shown to be especially effective in tasks that involve deictic referencing and guiding physical objects [Moll and

Sallnas, 2009].

We draw upon these insights to propose a formalism [2] for combining verbal communication and task actions, in order to guide a human teammate towards a better way of doing a task. We investigate the effect of different types of verbal communication in team performance and trust in the robot.

[2] Stefanos Nikolaidis, Minae Kwon, Jodi Forlizzi, and Siddhartha Srinivasa. Planning with verbal communication for human-robot collaboration. *Journal of Human-Robot Interaction (JHRI)*, 2018. (under review)

2.3.2 *Human-Robot Teams*

Verbal communication in human-robot teams has been shown to affect collaboration, as well as people's perception of the robot [Mavridis, 2015, Thomaz et al., 2016, Grigore et al., 2016]. Robot dialog systems have mostly supported human-initiated or robot-initiated communication in the form of requests. An important challenge for generating legible verbal commands has been symbol grounding [Mavridis, 2015, Tellex et al., 2011], which is described as the ability of the robot to map a symbol to a physical object in the world. Tellex et al. [2011] presented a model for inferring plans from natural language commands; inverting the model enables a robot to recover from failures, by communicating the need for help to a human partner using natural language [Tellex et al., 2014]. Khan et al. [2009] proposed a method for generating the minimal sufficient explanation that explains the policy of a Markov decision process, and Wang et al. [2016b] proposed generating explanations about the robot's confidence on its own beliefs. Recent work by Hayes and Shah [2017] has generalized the generation of explanations of the robot policies to a variety of robot controllers.

Of particular relevance is previous work in the autonomous driving domain [Koo et al., 2015]. Messages that conveyed "how" information, such as "the car is breaking," led to poor driving performance, whereas messages containing "why" information, such as "There is an obstacle ahead," were preferred and improved performance. Contrary to the driving domain, in our setting the human cannot verify the truthfulness of the robot "why" action. Additionally, unlike driving, in a physical human-robot collaboration setting there is not a clearly right action that the robot should take, which brings the human to a state of uncertainty and disagreement with the robot. In agreement with Koo et al. [2015], our results show the importance of finding the right away to explain robot behavior to human teammates.

Our work is also relevant to the work by Clair and Mataric [2015]. The authors explored communication in a shared-location collaborative task, using three different types of verbal feedback: self-narrative (e.g., "I'll take care of X"), role-allocative (e.g., "you handle X") and

empathetic (e.g., "Oh no" or "Great"). They showed that feedback improves both objective and subjective metrics of team performance. In fact, the robot's *verbal commands* ("Let's rotate the table clockwise") and *state-conveying actions* ("I think I know the best way of doing the task,") of our work resemble the role-allocative and self-narrative feedback. Additionally, Oudah et al. [2015] integrated verbal feedback about past actions and future plans into a learning algorithm, resulting in improved human-robot team performance in two game scenarios.

Contrary to existing work[3], our formalism enables the robot to reason about the effects of various types of verbal communication on the future actions of different human collaborators, based on their *internal state*. The human internal state captures inter-individual variability. Integrating it as a latent variable in a partially observable stochastic process allows the robot to infer online the internal state of a new human collaborator and decide when it is optimal to give feedback, as well as which type of feedback to give.

[3] In Devin and Alami [2016], the robot reasons over the human mental state, which represents the human knowledge of the world state and of the task goals. The human mental state is assumed to be fully observable by the robot.

3
Problem Formulation

Human-robot collaboration can be formulated as a *two player game with partial information*. We let x_t^{w} be the world state that captures the information that human and robot use at time t to take actions a_t^{R}, a_t^{H} in a collaborative task. Over the course of a task of total time duration T, robot and human receive an accumulated reward:

$$\sum_{t=1}^{T} R^{\mathrm{R}}(x_t^{\mathrm{w}}, a_t^{\mathrm{R}}, a_t^{\mathrm{H}})$$

for the robot and

$$\sum_{t=1}^{T} R^{\mathrm{H}}(x_t^{\mathrm{w}}, a_t^{\mathrm{R}}, a_t^{\mathrm{H}})$$

for the human.

We assume a robot policy π^{R}, which maps world states to actions. The human chooses their own actions based on a human policy π^{H}. If the robot could control both its own and the human actions, it would simply compute the policies that maximize its own reward.

However, the human is not another actuator that the robot can control. Instead, the robot can only *estimate* the human decision making process from observation and *make predictions* about future human behavior, which in turn will affect the reward that the robot will receive.

Therefore, the optimal policy for the robot is computed by taking the expectation over human policies π^{H}.

$$\pi^{\mathrm{R}^*} \in \arg\max_{\pi^{\mathrm{R}}} \mathbb{E}\left[\sum_{t=1}^{T} R^{\mathrm{R}}(x_t^{\mathrm{w}}, a_t^{\mathrm{R}}, a_t^{\mathrm{H}}) | \pi^{\mathrm{R}}, \pi^{\mathrm{H}}\right] \tag{3.1}$$

Solving this optimization is challenging: First, the human reward R^{H} may be unknown to the robot in advance. Second, even if the robot knows R^{H}, it may be unable to predict accurately the human actions, since human behavior is characterized by bounded

rationality [Kahneman, 2003]. Third, even if the human acts always rationally, exhaustively searching for the equilibria is computationally intractable in most cases [Papadimitriou, 2007]. Finally, even if $R^{\mathrm{H}} \equiv R^{\mathrm{R}}$, most real-world problems have multiple equilibria, and in the absence of a signaling mechanism, it is impossible to know which ones the agents will choose.

Therefore, rather than solving the game for the equilibria strategies, we make different assumptions about the human behavior within this general formulation. In the next chapters, we articulate these assumptions, and we explain how these lead to exciting and diverse robot behaviors in real-time interactions with actual human subjects, in manufacturing, personal robotics and assistive care settings.

4
Robot Adaptation

[1] In several manufacturing applications, such as assembly tasks, although important concepts such as tolerances and completion times are well-defined, many of the details are largely left up to the individualized preference of the mechanics. A robotic assistant interacting with a new human worker should be able to learn the preferences of its human partner in order to be an effective teammate. We assume a *leader-follower* teamwork model, where the human leader's preference is captured by the human reward function R^{H} and the human policy π^{H}. In this model, the goal of the robot is to execute actions aligned with the human preference. Therefore, in eq. 3.1 of chapter 3 we have:

$$R^{\mathrm{R}} \equiv R^{\mathrm{H}}$$

Learning of a Human Model. Learning jointly π^{H} and R^{H} can be challenging in settings where human and robot take actions simultaneously, and do not have identical action sets. To enable a robot to learn the human preference in collaborative settings, we looked at how humans communicate effectively their preferences in human teams. In previous work [Shah et al., 2011], insights from human teamwork have informed the design of a robot plan execution system which improved human-robot team performance. We focused on a team training technique known as *cross-training*, where human team-members switch roles to develop shared expectations on the task. This, in turn allows them to anticipate one another's needs and coordinate effectively. Using this insight, we proposed human-robot cross-training [Nikolaidis et al., 2015a, Nikolaidis and Shah, 2013, Nikolaidis et al., 2013], a framework where the robot learns a model of its human counter-part through two phases: a forward-phase, where human and robot follow their pre-defined roles, and a rotation phase, where the roles of human and robot are switched. The forward phase enables the robot to observe the human actions and estimate the human policy π^{H}. The rotation phase allows the robot

[1] This chapter summarizes for completion the work done in collaboration with Keren Gu, Premyszlaw Lasota, Ramya Ramakrishnan and Julie Shah, presented in [Nikolaidis, 2014].

Figure 4.1: Cross-training in a virtual environment leads to fluent human-robot teaming.

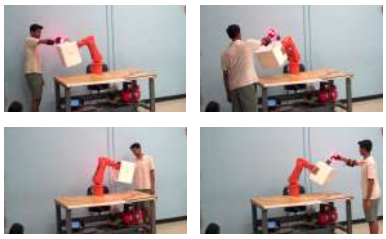


Figure 4.2: We clustered participants into types, based on how their preference of executing a hand-finishing task with the robot.

to observe the human inputs on the robot actions and infer the human preference R^{H} from human demonstrations [Argall et al., 2009]. After each training round, which occurs in a virtual environment, the robot uses the new estimates to compute the optimal policy from eq. 3.1. Our studies showed that cross-training provides significant improvements in quantitative metrics of team fluency, as well as in the perceived robot performance and trust in the robot (fig. 4.1). These results provide the first indication that effective and fluent human-robot teaming may be best achieved by modeling effective training practices for human teamwork.

Inference of a Human Type. Cross-training works by learning an individualized model for each human teammate. For more complex tasks, this results in a large number of training rounds, which can be tedious from a human-robot interaction perspective. However, our pilot studies showed that even when there was a very large number of feasible action sequences towards task completion, people followed a limited number of "dominant" strategies. Using this insight, we used unsupervised learning techniques to identify distinct *human types* from joint-action demonstrations (fig. 4.2) [Nikolaidis et al., 2015b]. For each type $\theta \in \Theta$, we used supervised learning techniques to learn the human reward $R^{\mathrm{H}}(x_t^{\mathrm{w}}, a_t^{\mathrm{R}}, a_t^{\mathrm{H}}; \theta)$, as well as the human policy $\pi^{\mathrm{H}}(x_t^{\mathrm{w}}; \theta)$, This simplified the problem of learning $R^{\mathrm{H}}, \pi^{\mathrm{H}}$ of a new human worker, to simply inferring their type θ. We enabled this inference by denoting the human type as a latent variable in a partially observable stochastic process (POMDP). This allowed the robot to take information seeking actions in order to infer online the type a new user, and execute actions aligned with the preference of that type. This draws upon insights from previous work on cooperative games [Macindoe et al., 2012] and vehicle navigation [Bandyopadhyay et al., 2013], where the human intent was modeled as a latent variable in a POMDP, albeit with prespecified models of human types. In a human subject experiment, participants found that the robot executing the computed policy anticipated their actions, and in complex robot configurations they completed the task faster than

manually annotating robot actions.

5
Human Adaptation

As robotics systems become more advanced, they have access to information that the human may not have; this suggests that rather than always following the human, they could use this information as *equal partners* to guide their human teammates towards better ways of doing the task. In that case, it is no longer optimal for the robot to optimize the human reward function; instead, the robot should maximize its own reward function, which is different than the human's:

$$R^{\mathrm{R}} \neq R^{\mathrm{H}}$$

An improvement upon the leader-follower setting is to recognize that the human policy can change based on the robot actions. We let a history of world states, human and robot actions h_t:

$$h_t = (x_0^{\mathrm{w}}, a_0^{\mathrm{R}}, a_0^{\mathrm{H}}, ..., x_t^{\mathrm{w}}, a_t^{\mathrm{R}}, a_t^{\mathrm{H}})$$

Given this history, the human policy $\pi^{\mathrm{H}}(x_t^{\mathrm{w}}, h_t; \theta_t)$ is a function not only of the current world state x_t^{w} and human type θ_t, but also of the history h_t. Modeling the human policy as a function of the robot actions and solving the optimization of eq. 3.1, chapter 3, makes the interaction an *underactuated dynamical system*, where the robot reasons over how its own actions affect future human actions, and takes that into account into its own decision making.

5.1 A Bounded Memory Model.

This history h_t can grow arbitrarily large, making optimizing for the robot actions computationally intractable. In practice, however, people do not have perfect recall. Using insights from work on bounded rationality in behavioral economics, we simplify the optimization, using a Bounded memory human Adaptation Model (BAM) [1].

The Bounded memory human Adaptation Model specifies a parameterization of the human policy π^{H}. We define a set of *modal*

Work done in collaboration with David Hsu.

[1] Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddhartha Srinivasa. Formalizing human-robot mutual adaptation: A bounded memory model. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016; and Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research (IJRR)*, 2017a

policies or *modes* M, where $m \in M$ is a stochastic policy mapping states and histories to joint human-robot actions: $m : X^{w} \times H_t \rightarrow \Pi(A^{R}) \times \Pi(A^{H})$.

At each time-step, the human has a mode $m^{H} \in M$ and perceives the robot as following a mode $m^{R} \in M$. Then in the next time-step the human may switch to m^{R} with some probability α. If m^{H} maximizes the expected accumulated reward, the robot optimal policy would be to follow m^{R}, expeting the human to adapt.

Specifically, we model the human policy π^{H} as a probabilistic finite-state automaton (PFA), with a set of states $Q : X^{w} \times H_t$. A joint human-robot action a^{H}, a^{R} triggers an emission of a human and robot modal policy $f : Q \rightarrow \Pi(M) \times \Pi(M)$, as well as a transition to a new state $P : Q \rightarrow \Pi(Q)$.

5.1.1 *Bounded Memory Assumption*

Herbert Simon proposed that people often do not have the time and cognitive capabilities to make perfectly rational decisions, in what he described as “bounded rationality" [Simon, 1979]. This idea has been supported by studies in psychology and economics [Kahneman, 2003]. In game theory, bounded rationality has been modeled by assuming that players have a “bounded memory” or “bounded recall" and base their decisions on recent observations [Powers and Shoham, 2005, Monte, 2014, Aumann and Sorin, 1989]. In this work, we introduce the bounded memory assumption in a human-robot collaboration setting. Under this assumption, humans will choose their action based on a history of k-steps in the past, so that $Q : X^{w} \times H_k$.

5.1.2 *Fully Observable Modal Policies*

This section proposes a method for inference of m^{H} and m^{R}, when the modes are fully observable. The general case of partially observable modes is examined in chapter 6, section 6.2.1.

If the modes are fully observable, it is sufficient to retain only the k-length mode history, rather than h_k, simplifying the problem. We define a set of features, so that $\phi(q) = \{\phi_1(q), \phi_2(q), ..., \phi_N(q)\}$. We can choose as features the frequency counts ϕ^{H}_{μ}, ϕ^{R}_{μ} of the modal policies followed in the interaction history, so that:

$$\phi^{H}_{\mu} = \sum_{i=1}^{k} [\mu^{H}_{i} = \mu] \quad \phi^{R}_{\mu} = \sum_{i=1}^{k} [\mu^{R}_{i} = \mu] \;\; \forall \mu \in M \tag{5.1}$$

μ^{H}_{i} and μ^{R}_{i} is the modal policy of the human and the robot i time-steps in the past. We note that k defines the history length, with $k = 1$ implying that the human will act based only on the previous interaction. Drawing upon insights from previous work which assumes

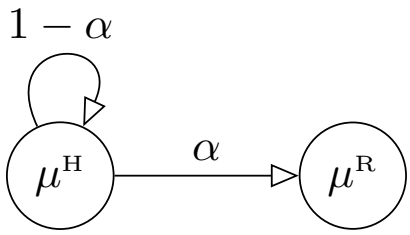


Figure 5.1: The BAM human adaptation model.

maximum likelihood observations for policy computation in belief-space [Platt et al., 2010], we used as features the modal policies with the maximum frequency count:

$$\mu^{\mathrm{H}} = \arg\max_{\mu} \phi^{\mathrm{H}}_{\mu}, \quad \mu^{\mathrm{R}} = \arg\max_{\mu} \phi^{\mathrm{R}}_{\mu} \tag{5.2}$$

The proposed model does not require a specific feature representation. For instance, we could construct features by combining modal policies μ^{H}_i, μ^{R}_i using an arbitration function [Dragan and Srinivasa, 2012].

5.1.3 *Human Adaptability*

We define the adaptability as the probability of the human switching from their mode to the robot mode. It would be unrealistic to assume that all users are equally likely to adapt to the robot. Instead, we account for individual differences by parameterizing the transition function P by the *adaptability* α of an individual. Then, at state q the human will transition to a new state by choosing an action specified by μ^{R} with probability α, or an action specified by μ^{H} with probability $1-\alpha$ (fig. 5.1).

In order to account for unexpected human behavior, we assign uniformly a small, non-zero probability ϵ for the human taking a random action of some mode other than μ^{R}, μ^{H}. The parameter ϵ plays the role of probability smoothing. In the time-step that this occurs, the robot belief on α will not change. In the next time-step, the robot will include the previous human action in its inference of the human mode μ^{H}.

We note that the Finite State Machine in fig. 5.1 shows the human mode transition in one time-step only. For instance, if the human switches from μ^{H} to μ^{R} and $k = 1$, in the next time-step the new human mode μ^{H} will be what was previously μ^{R}. In that case, oscillation between μ^{R} and μ^{H} can occur. We discuss this in section 6.1.3.3.

Throughout this chapter, we assume that the adaptability known to the robot and fixed throughout the task. We relax the first assumption in chapter 6 and the second assumption in chapter 7.

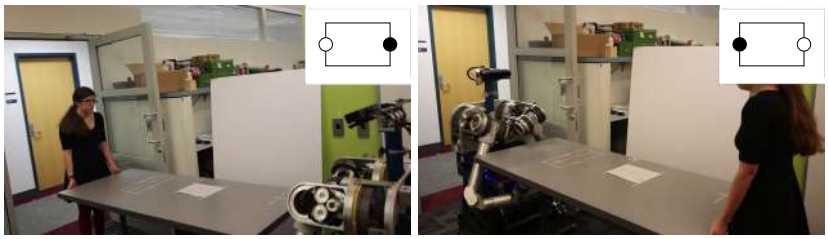


Figure 5.2: A human and a robot collaborate to carry a table through a door. (left) The robot prefers facing the door (Goal A), as it has a full view of the door. (right) The robot faces away from the door (Goal B).

5.1.4 *Characterizing Modal Policies*

At each time-step, the human and robot modes are not directly observed, but must be inferred from the human and robot actions. This can be achieved by characterizing a set of modal policies through one of the following ways:

Manual specification In some cases the modal policies can be easily specified. For instance, if two agents are crossing a corridor (Section 6.1.5), there are two deterministic policies leading to task completion, one for each side. Therefore, we can infer a mode directly from the action taken.

Learning from demonstration In previous work, joint-action demonstrations on a human-robot collaborative task were clustered into groups and a reward function was learned for each cluster [Nikolaidis et al., 2015b], which we can then associate with a mode.

Planning-based prediction Previous work assumes that people move efficiently to reach destinations by optimizing a cost-function, similarly to a goal-based planner [Ziebart et al., 2009]. Given a set of goal-states and a partial trajectory, we can associate modes with predictive models of future actions towards the most likely goal.

Computation of Nash Equilibria Following a game-theoretic approach, we solve the stochastic game described in chapter 3 and restrict the set of modal policies to the equilibrium strategies. For instance, we can formulate the example of human and robot crossing a corridor as a coordination game, where strategies of both agents moving on opposite sides strictly dominate strategies where they collide.

5.1.5 *Application*

We show the applicability of the model in an example table-carrying task (fig. 5.2), A human and HERB [Srinivasa et al., 2010], an autonomous mobile manipulator, work together to carry a table out of the room. There are two strategies: the robot facing the door (Goal A) or the robot facing away from the door (Goal B). We assume that the robot prefers Goal A, as the robot's forward-facing sensor has a clear

view of the door, leading to better task performance. Not aware of this, an inexperienced human partner may prefer Goal B. If human and robot rotate the table in the same direction, the table orientation changes at a small amount. Otherwise, the table does not move. We assume two deterministic and fully observable modal policies, one for each goal. We will show that the Bounded memory Adaptation Model allows the robot to reason over the probability of the human changing its future behavior and to choose its own actions in return.

Robot Policy Computation. We define an infinite-horizon Markov decision process [Russell and Norvig, 2003] in this setting as a tuple $\{X, A^{\mathrm{R}}, P, R^{\mathrm{R}}, \gamma\}$, where:

- $X : X^{\mathrm{w}} \times \Theta \times H_k$ is the set of observable states. X^{w} is the set of world states, Θ is the set of human states and H_k the set of recent histories. The world states are the different table configurations. The set of human states is defined as $\Theta : M \times \mathcal{A}$. A human state is the vector $\theta = (m^{\mathrm{H}}, \alpha)$, where m^{H} is the human mode and α the human adaptability.
- A^{R} is a finite set of robot actions. A robot action is a discrete rotation of the table.
- $P : X \times A^{\mathrm{R}} \longrightarrow \Pi(X)$ is the state transition function, indicating the probability of reaching a new state x' from state x and action a^{R}. Given a table configuration x^{w}, and a table rotation a^{R}, the next table configuration depends on the human action. The probability of the human action is given by the BAM model, and it is a function of their mode, the current world state, the history of interactions and their adaptability.
- $R^{\mathrm{R}} : X^{\mathrm{w}} \times H_k \longrightarrow \mathbb{R}$ is the reward function, giving the immediate reward that the human-robot receives. We assume a set of goal states G, which in the table-carrying example are the two table configurations of fig. 5.2. We specify the reward function as follows

$$R(x^{\mathrm{w}}, h_k) = \begin{cases} R_{\mathrm{goal}} > 0 & : x^{\mathrm{w}} \in G \\ C < 0 & : x^{\mathrm{w}} \notin G \text{ and } m^{\mathrm{R}} \neq m^{\mathrm{H}} \\ 0 & : \text{otherwise} \end{cases} \tag{5.3}$$

  There is a positive reward R_{goal} associated with each goal, and a negative cost C associated with human-robot mode disagreement. We assume that the modes m^{R} and m^{H} are inferred from the history h_k, as explained in section 5.1.2.
- γ is a discount factor. The discount factor implicitly penalizes disagreement, since when human and robot disagree the table does not move and the expected reward decreases.

The problem of finding the optimal policy of eq. 3.1 is reduced to solving the above MDP. We can do this using dynamic programming.

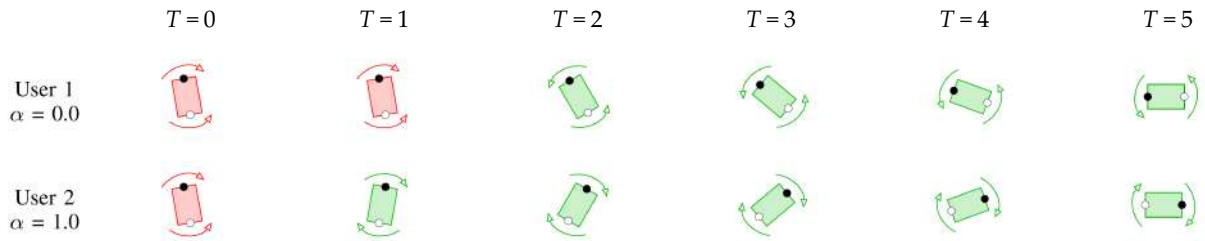


Figure 5.3: Sample runs on the human-robot table-carrying task, with two simulated humans of adaptability level α=0 and α=1. A fully adaptable human has α=1, while a fully non-adaptable human has α=0. Red color indicates human (white dot) and robot (black dot) disagreement in their actions, in which case the table does not move. User 1 is non-adaptable, and the robot complies. User 2 is adaptable, and the robot successfully guides them towards a better strategy.

Work done in collaboration with Ariel Procaccia and Swaprava Nath.

Interestingly, if R_{goalA} is much larger than R_{goalB} and C is negligible, the robot will always insist towards the optimal goal, ignoring the user. On the other hand, if the cost of disagreement C is very high, the robot will always adapt to the user. For some appropriate values of R_{goalA}, R_{goalB} and C, the robot will choose its actions based on the user adaptability (fig. 5.3). If the user is adaptable, the robot will insist towards the optimal goal, expecting that the user will change their actions in the future. On the other hand, if the user is non-adaptable, the robot expects them to keep disagreeing and it will change its own actions instead. This behavior matches our intuition. In chapter 6 we show that this very capability enables the robot to guide users towards better ways of completing the task, while retaining their trust in the robot.

5.2 *A Best-Response Model.*

A particular instance of treating interaction as an underactuated system is modeling people as computing a *best-response* to the last robot action using their reward function R^{H}:

$$\pi^{\text{H}}(x_t^{\text{w}}, a_t^{\text{R}}; \theta_t) \in \arg\max_{a_t^{\text{H}}} R^{\text{H}}(x_t^{\text{w}}, a_t^{\text{R}}, a_t^{\text{H}}; \theta_t) \tag{5.4}$$

This draws upon insights from previous work on a particular class of Stackelberg games [Conitzer and Sandholm, 2006], the *repeated Stackelberg security games* [Balcan et al., 2015]. In this setting, the follower observes the leader's possibly randomized strategy, and chooses a best-response. We extend this model to a human-robot collaboration setting, where the leader is the robot and the follower is the human, and we model human adaptation by having the follower's reward stochastically changing over time [2].

The change in the reward occurs, as the human observes the outcomes of the robot and their own actions and updates its expectations on the robot's capabilities. This model allows the robot to *reason over how the human expectations of the robot capabilities will change based*

[2] Stefanos Nikolaidis, Swaprava Nath, Ariel D Procaccia, and Siddhartha Srinivasa. Game-theoretic modeling of human adaptation in human-robot collaboration. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017b

on its own actions. Computing an optimal policy with this model enables the robot to decide optimally between *communicating information* to the human and *choosing the best action given the information that the human currently has.*

We prove that, if the robot can observe whether the user has learned at each round, the computation of the optimal policy is simple (Lemmas 1 and 2), and can be done in time polynomial in the number of robot actions and the number of rounds (Theorem 1).

We show through a human subject experiment in a table-clearing task that the proposed model significantly improves human-robot team performance, compared to policies that assume complete human adaptation to the robot. Additionally, we show through simulations that the proposed model performs well for a variety of randomly generated tasks. This is the first step towards modeling the change of human expectations of the robot capabilities through interaction, and integrating the model into robot decision making in a principled way.

5.2.1 Formulation

We follow the two-player game formulation of chapter 3. Human and robot have a *finite* set of robot and human actions denoted by $A^{\mathrm{R}} = \{a_1^{\mathrm{R}}, \ldots, a_m^{\mathrm{R}}\}$ and $A^{\mathrm{H}} = \{a_1^{\mathrm{H}}, \ldots, a_n^{\mathrm{H}}\}$. We make the additional assumption of a repeated game, with only one world state, i.e., $|X^{\mathrm{w}}| = 1$.

The payoff [3] associated with each pair of actions is uniquely identified by the robot reward $R^{\mathrm{R}} = [r_{i,j}], (i,j) \in [m] \times [n]$, where the entry $r_{i,j}$ denotes the reward for the action pair $(a_i^{\mathrm{R}}, a_j^{\mathrm{H}})$ chosen by these two players. We denote the reward vector corresponding to row i by r_i, i.e., $r_i = (r_{i,1}, \ldots, r_{i,n})$. Importantly, the *same reward* is experienced together by both players. Therefore this is an *identical payoff* game where the goal is to maximize the total reward obtained in T (finite) rounds of playing this repeated game. If the reward matrix was perfectly known to both the agents, they would have played the action pair that gives the maximum reward in each round.

[3] We will use the terms 'reward' and 'payoff' interchangeably.

However, we assume that in the beginning of the game, the robot has perfect information about the reward matrix, whereas the human has possibly incorrect information (captured by a reward matrix R^{H} which the human *believes* to be the true reward matrix). In different rounds of the game, the human probabilistically learns different entries of this matrix and picks action accordingly. We will assume that the human is capable of taking the optimal action given their knowledge of the payoffs, e.g., if a specific row of this matrix is completely known to the human and the robot plays the action corresponding to

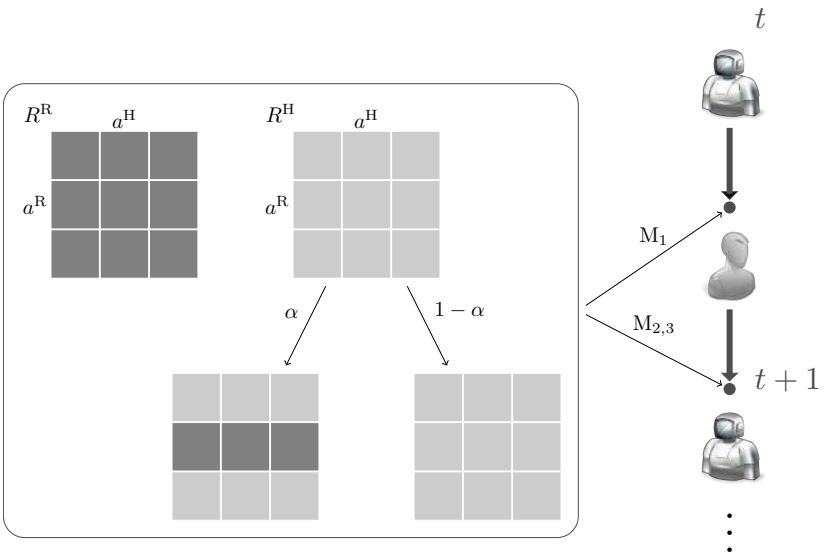


Figure 5.4: Models of human partial adaptation, described in section 5.2.2. The human learns with probability α the entries of row r_i that correspond to the robot action a_i^{R} played, and with probability 1-α none of the entries. The learning occurs before her action (*learning from robot action* – $\mathbf{M_1}$), or after her action (*learning from experience* – full observability ($\mathbf{M_2}$) or partial observability ($\mathbf{M_3}$)).

this row [4], the human will pick the action that maximizes the reward in this row. However, if the entries of a row are yet to be learned by the human, the human picks an action according to $\arg\max r_i^{\mathrm{H}}$, where r_i^{H} is the i-th row of R^{H}.

The only aspect of this game that may change over time is the state of the human, which we denote by θ_t, $t \in [T]$. Therefore, the state of the game is simply the state of the human agent. We denote the state space of the game as Θ; it will be instantiated below in different models of information dissemination.

With these assumptions, we can simplify eq. 3.1 from chapter 3 as follows:

$$\pi^{\mathrm{R}^*} \in \arg\max_{\pi^{\mathrm{R}}} \mathbb{E}\left[\sum_{t=1}^{T} R^{\mathrm{R}}(a_t^{\mathrm{R}}, a_t^{\mathrm{H}}; \theta_t) | \pi^{\mathrm{R}}, \pi^{\mathrm{H}}\right] \tag{5.5}$$

From eq. 5.4, the human policy π^{H} is deterministic; the expectation is taken only with respect to the future human states θ_t.

5.2.2 *Approach*

We consider a setting where, in each round, the robot plays first by choosing a row. We model the strategy of the human $\pi^{\mathrm{H}} : A^{\mathrm{R}} \times \Theta \to A^{\mathrm{H}}$ as maximizing a human reward function R^{H}. In other words, the human *best responds* to the robot action, according to the (possibly erroneous) way she currently perceives the payoffs. The human reward matrix R^{H} evolves over time, as the human learns the "true" reward R^{R} through interaction with the robot. We propose a model of human

[4] We will refer to this robot action as *playing a row*.

partial adaptation, where the human learns with probability α the entries of row r_i that correspond to the robot action a_i^R played, and with probability $(1-\alpha)$ none of the entries. We consider the following models, based on when the human learning occurs, and on whether the robot directly observes if the human has learned.

M_1. The human learns the payoffs immediately after the robot plays a row, and before she takes her own action. The robot can infer whether the human has learned the row, by observing the reward after the human has played in the same round. We call this *learning from robot action*, where the robot has *full observability* of the human internal state. This model is studied in section 5.2.4.1.

M_2. The human learns the payoffs associated with a row after she plays in response to the robot's action. The robot can observe whether the human has learned before the start of the next round, for instance by directly asking the human, or by interpreting human facial expressions and head gestures [El Kaliouby and Robinson, 2005]. We call this model *learning from experience*, where the robot has *full observability* of the human internal state. This model is studied in section 5.2.4.2.

M_3. Identically to model **M_2**, the human learns a row after her action in response to the robot action. However, the robot does not immediately observe whether the human has learned, rather infers it through the observation of human actions in subsequent rounds of the game. This is a case of *learning from experience*, *partial observability*.

We note that we do not define a model for *learning from robot action*, *partial observability* case, since the robot can always directly observe whether the human has learned, based on the reward resulting from the human action in the same round.

Figure 5.4 shows the different models. In Section 5.2.3, we discuss the general case of partial observability (Model **M_3**) and formulate the problem as a Markov Decision Process [Russell and Norvig, 2003]. Computing the optimal policy in this case is exponential in the number of robot actions m. However, when the robot has full observability of the human state (Models **M_1**, **M_2**), the optimal policy has a special structure and can be computed in time polynomial in m and T (Section 5.2.4).

5.2.3 *Theory: Partial Observability*

In this section we examine the hardest case, where the human learns the payoffs associated with the row after their choice of actions, and the robot cannot directly observe whether the human has learned the payoffs (model **M_3**). Instead, the robot infers whether the human

has learned the row by observing the human response in subsequent rounds of the game.

While the human state is partially observable, we can exploit the structure of the problem and reduce it to a Markov Decision Process based on the following observation: the probability of the human having learned a row is either 0 when it is played for the first time; α after it is played by the robot and the human responds sub-optimally; and 1 after the the human has played the actual best-response strategy (according to R) for that row (which means she has learned the true rewards in the previous round).

We define a Markov decision process in this setting as a tuple $\{X, A^{\mathrm{R}}, P, R, T\}$, where:

- $X \in \{0, \psi, 1\}^m$ is a finite set of human states (so that $X \equiv \Theta$). A human state x is represented by a vector $(x_1, x_2, ..., x_m)$, where $x_i \in \{0, \psi, 1\}$ and i is the corresponding row in the matrix. The starting state is $x_i = 0$ for each row i. $x_i = \psi$ indicates that the robot does not know whether human has learned row i or not. In this state, the human plays the best response in that row with probability α, or an action defined by the strategy π^{H} of the human with probability $(1 - \alpha)$ [5]. If the human plays best-response, then the robot knows that human has learned row i, thus the entry for that row is $x_i = 1$.
- $A^{\mathrm{R}} = \{a_1^{\mathrm{R}}, \ldots, a_m^{\mathrm{R}}\}$ is a finite set of robot actions.
- $P : X \times A^{\mathrm{R}} \longrightarrow \Pi(X)$ is the state transition function, indicating the probability of reaching a new state x' from state x and action a_i^{R}. State x transitions to a new state x' with all vector entries identical, apart from the element x_i corresponding to the row played. If the robot plays i for the first time ($x_i = 0$), the corresponding entry in the next state x' deterministically becomes $x_i' = \psi$, since the robot no longer knows whether the human has learned the payoffs for that row. If $x_i = \psi$, the human may have learned that row in the past and play the best-response strategy, leading to a transition to $x_i' = 1$ with probability α. If the human does not play the best-response strategy, the robot still does not know whether they will have learned the payoffs after the current round, thus $x_i' = \psi$ with probability $(1 - \alpha)$. If $x_i = 1$, the corresponding entry in all subsequent states will be $x_i' = 1$, i.e., if the human learns a row, we assume that they remember the row in the future.
- $R : A^{\mathrm{R}} \times A^{\mathrm{H}} \longrightarrow \mathbb{R}$ is the reward function, giving the immediate reward gained by performing a human and robot action. Note that if action i is played and the state has $x_i = \psi$, the reward will be based on the best response in row i of R with probability α, and on row i of R^{H} with probability $(1 - \alpha)$ — we consider the *expected* reward. We assume that the robot knows the "true" reward, so that

[5] We assume that α is a parameter known to the robot and fixed throughout the task

$R^{\mathrm{R}} \equiv R$.

- T is the number of rounds.

The robot's decision problem is to find the optimal policy $\pi^{\mathrm{R}^*} = (\pi_1^{\mathrm{R}^*}, \ldots, \pi_T^{\mathrm{R}^*})$ to maximize the expected payoff, as defined in eq. 5.5.

We observe that in the current formalism, the size of the state-space is $|X| = 3^m$, where m is the number of robot actions. Therefore, the computation of the optimal policy requires time exponential in m. In Section 5.2.4, we show that for the case where the robot can observe whether the human has learned the payoffs, the optimal policy can be computed in time polynomial in m and T.

5.2.4 *Theory: Full Observability*

In this section, we assume that the robot can observe whether the human has learned the payoffs. We instantiate state x_t as a vector $(x_{t,1}, x_{t,2}, \ldots, x_{t,m})$, where each $x_{t,i}$ is now a binary variable in $\{0, 1\}$ denoting the robot's knowledge in round t of whether row i is learned by the human. In contrast to section 5.2.3, there is no uncertainty about whether the human has learned or not (therefore no ψ state).

5.2.4.1 Learning from Robot Action

This is the scenario where the human might learn the payoffs immediately after the robot plays a row, and before she takes their own action (Model $\mathbf{M_1}$ in section 5.2.2). Clearly, the robot can figure out if the human learned the row by observing the reward for that round. Our algorithmic results in this model strongly rely on the following lemma.

Lemma 1. *In model* $\mathbf{M_1}$*, if, under the optimal policy* π^{R^*}*, there exists* $\tau \in \{2, \ldots, T\}$ *and* $i \in [m]$ *such that* $x_{\tau,i} = 1$ *and* $\max r_i \geqslant \max r_j$ *for all* j *such that* $x_{\tau,j} = 1$*, then* $\pi_t^{R^*}(x_t) = a_i^R$ *for all* $\tau \leqslant t \leqslant T$ *and for all* $x_t = x_\tau$.

This lemma says that the optimal policy for the robot is to pick the action a_i^R when i is the row that yields the maximum reward among the rows already learned by the human. As we will show in detail later, this directly leads to a computationally efficient algorithm, via the following insight: *if the robot plays a row and this row is successfully revealed to the human, the optimal policy for the robot is to keep playing that row until the end of the game.*

The main idea behind the proof below is [6]: if at round $t - 1$ the optimal policy plays row 2, and that row is revealed, then it will not explore the unrevealed (higher rewarding) row 1 afterwards. The reason is that if the optimal policy chose to explore row 1 at some

[6] Informally, one way to understand why the lemma holds is by thinking that, if the robot chooses between a high-cost high-reward and a low-cost low-reward action, it is better to choose the high-cost high-reward action as early as possible, so that it has enough time to reap the benefits if the human succeeds in learning.

time in the future — which is a contradiction to the lemma — then playing row 1 at round $t-1$ would have been optimal, therefore an optimal policy would not have played row 2 at round $t-1$.

Proof of Lemma 1. Assume for contradiction that the lemma does not hold, and let t be the *last* round in which the optimal policy violates the lemma, i.e., the last round in which there are $i, j \in [m]$ such that $x_{t,i} = 0$ and $x_{t,j} = 1$, but the optimal policy plays row i. Without loss of generality assume that these i and j are rows 1 and 2, respectively. For all rounds from $t+1$ to T, it holds (by the choice of t) that if row i is revealed to the human, the optimal policy will continue playing a_i^{R} (if there are multiple such rows, it plays the one with highest reward).

Let the maximum rewards corresponding to rows 1 and 2 be R_1 and R_2, respectively, i.e., $R_k = \max r_k$. We assume w.l.o.g. that row 2 has the highest maximum reward among all revealed rows. We can also assume that $R_1 > R_2$, since a policy that moves away from a row that is simultaneously known and more rewarding is clearly suboptimal.

If a row is not learned, the reward associated with actions a_1^{R} and a_2^{R} are C_1 and C_2, where $C_k = r_k[\operatorname{argmax} r_k^H]$. Clearly, $C_1 \leqslant R_1$ and $C_2 \leqslant R_2$. We define U_1, so that:

$$U_1(\pi^{\mathrm{R}}|x_1) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} R(\pi_t^{\mathrm{R}}(x_t), \pi^{\mathrm{H}}(\pi_t^{\mathrm{R}}(x_t), x_t)) \,\middle|\, x_1\right] \tag{5.6}$$

Since the optimal policy chose a_1^{R} in round t over a_2^{R}, the expected payoff of choosing a_1^{R} in round t must be larger than that of a_2^{R}, i.e.,

$$\begin{aligned}&\alpha(R_1 + U_{t+1}(\pi^{\mathrm{R}^*}|(1,1,\ldots))) + (1-\alpha)\cdot(C_1 + U_{t+1}(\pi^{\mathrm{R}^*}|(0,1,\ldots)))\\ &\quad > R_2 + U_{t+1}(\pi^{\mathrm{R}^*}|(0,1,\ldots)),\end{aligned}$$

where the first term on the LHS shows the expected payoff if row 1 is learned in round t, and the second term shows the payoff when it is not. It follows that

$$\begin{aligned}&\alpha(R_1 + R_1\cdot(T-t-1)) + (1-\alpha)\cdot(C_1 + R_2\cdot(T-t-1))\\ &\quad > R_2 + R_2\cdot(T-t-1).\end{aligned} \tag{5.7}$$

The implication holds because from round $t+1$, we assume (by the choice of t) that the optimal policy continues playing the best action among the revealed rows. We make the above inequality into an equality by adding a slack variable $\epsilon > 0$ as follows.

$$\begin{aligned}&\alpha R_1\cdot(T-t) + (1-\alpha)(C_1 + R_2\cdot(T-t-1))\\ &\quad = R_2 + R_2\cdot(T-t-1) + \epsilon.\end{aligned} \tag{5.8}$$

Denote the LHS of the above equality as ρ_1. Note that this is the assumed optimal value of the objective function at round t when the

state x_t is $(0, 1, \ldots)$, i.e., $U_t(\pi^{R^*}|(0, 1, \ldots)) = \rho_1$. Rearranging the expressions above, we get,

$$\alpha R_1 \cdot (T - t) + (1 - \alpha)C_1 = R_2 + \alpha R_2 \cdot (T - t - 1) + \epsilon. \tag{5.9}$$

We claim that if the optimal policy chooses the action a_1^R at round t, then the expected payoff in round $t - 1$ from choosing the action a_1^R would have been larger than that of the action a_2^R. If our claim is true, then the current policy, which chose a_2^R at $t - 1$, cannot be optimal, and we reach a contradiction. To analyze the decision problem in round $t - 1$, we need to consider two possible states of the game in this round.

Case 1: $x_{t-1} = (0, 0, \ldots)$. In this state, playing a_1^R gives an expected payoff of

$$\begin{aligned} &\alpha(R_1 + U_t(\pi^{R^*}|(1, 0, \ldots))) + (1 - \alpha)(C_1 + U_t(\pi^{R^*}|(0, 0, \ldots))) \\ &\quad \geqslant \alpha(R_1 + R_1(T - t)) + (1 - \alpha)(C_1 + U_t(\pi^{R^*}|(0, 0, \ldots))). \end{aligned} \tag{5.10}$$

The inequality holds because in state $(1, 0, \ldots)$, playing a_1^R yields at least R_1 in every subsequent round. Playing a_2^R in round $t - 1$ yields,

$$\alpha(R_2 + \rho_1) + (1 - \alpha)(C_2 + U_t(\pi^{R^*}|(0, 0, \ldots))). \tag{5.11}$$

This expression is similar to the RHS of Equation (5.10), except that the expected payoff at $x_t = (0, 1, \ldots)$ is assumed to be ρ_1. We claim that the expression on the RHS of eq. (5.10) is larger than the expression in eq. (5.11), for which we need to show that

$$\begin{aligned} &\alpha(R_1 + R_1(T - t)) + (1 - \alpha)C_1 \\ &\quad > \alpha(R_2 + \rho_1) + (1 - \alpha)C_2 \\ \iff &\alpha R_1 + R_2 + \alpha R_2 \cdot (T - t - 1) + \epsilon \\ &\quad > \alpha(R_2 + R_2 + R_2 \cdot (T - t - 1) + \epsilon) + (1 - \alpha)C_2 \\ \iff &\alpha R_1 + R_2 + \epsilon > \alpha R_2 + \alpha R_2 + (1 - \alpha)C_2 + \alpha\epsilon. \end{aligned}$$

In the first equivalence, we substitute the expression from eq. (5.9) on the LHS and the expression of ρ_1 from eq. (5.8) on the RHS. The second equivalence holds by canceling out one term. We see that the final inequality is true since $R_2 \geqslant C_2$, $R_1 > R_2$, and $0 < \alpha < 1$.[7]

[7] If $\alpha = 1$, playing the row $\arg\max R_i$ is optimal and the lemma holds trivially. For $\alpha = 0$, the lemma is vacuously true. So, we assume $0 < \alpha < 1$ w.l.o.g.

Case 2: $x_{t-1} = (0, 1, \ldots)$, in this state playing the action a_1^R gives an expected payoff of at least

$$\begin{aligned} &\alpha(R_1 + R_1 \cdot (T - t)) + (1 - \alpha)(C_1 + U_t(\pi^{R^*}|(0, 1, \ldots))) \\ &\quad = \alpha(R_1 + R_1 \cdot (T - t)) + (1 - \alpha)(C_1 + \rho_1). \end{aligned} \tag{5.12}$$

This is similar to the RHS of eq. (5.10) except that now we can replace $U_t(\pi^{R^*}|(0, 1, \ldots))$ with ρ_1. On the other hand, the expected payoff of

the action a_2^R in round $t-1$ is given by $R_2 + \rho_1$ — because at state $(0, 1, \ldots)$ in round $t-1$, action a_2^R gives R_2 deterministically, since the human knows row 2. The state remains the same even after reaching round t. The expected payoff at this round for this state is assumed to be ρ_1. Now to show that the expression in eq. (5.12) is larger than $R_2 + \rho_1$, we need to show that

$$\begin{aligned}
&\alpha(R_1 + R_1 \cdot (T-t)) + (1-\alpha)(C_1 + \rho_1) > R_2 + \rho_1 \\
\iff &\alpha R_1 + \alpha R_1 \cdot (T-t) + (1-\alpha)C_1 > R_2 + \alpha\rho_1 \\
\iff &\alpha R_1 + R_2 + \alpha R_2 \cdot (T-t-1) + \epsilon \\
&\qquad > R_2 + \alpha R_2 + \alpha R_2 \cdot (T-t-1) + \alpha\epsilon \\
\iff &\alpha R_1 + \epsilon > \alpha R_2 + \alpha\epsilon
\end{aligned}$$

The first equivalence comes from reorganizing the inequality. The second equivalence is obtained through substitution using eqs. (5.8) and (5.9). The third equivalence follows by canceling out two terms. The last inequality is true since $R_1 > R_2$ and $0 < \alpha < 1$.

To summarize, we have reached a contradiction in both cases, which are exhaustive. This proves the lemma. □

5.2.4.2 Learning from Experience

Recall that in model $\mathbf{M_2}$, the human learns with probability α all payoffs associated with a row *after* they play their action in response to the robot playing an unrevealed row. They do not learn with probability $1-\alpha$. This model is the same as model $\mathbf{M_3}$ of section 5.2.3, with an additional assumption: before the robot takes its next action, it can observe the current state.

We show that in this setting too, the optimal policy has a special structure similar to that under model $\mathbf{M_1}$ (section 5.2.4.1), which can be computed in time polynomial in m and T.

Lemma 2. *In model $\mathbf{M_2}$, if, under the optimal policy π^{R^*}, there are $\tau \in \{2, \ldots, T\}$ and $i \in [m]$ such that $x_{\tau,i} = 1$ and $\max r_i \geqslant \max r_j$ for all j such that $x_{\tau,j} = 1$, then $\pi_t^{R^*}(x_t) = a_i^R$ for all $\tau \leqslant t \leqslant T$ and for all $x_t = x_\tau$.*

The proof is similar to the proof of Lemma 1. However, the expected payoffs and the corresponding inequalities are different. Therefore, we provide a proof sketch that identifies the differences from the previous proof.

Proof of Lemma 2 (sketch). As before, the idea of the proof is to show that if the optimal policy changes its action from playing the revealed row that yields maximum reward, a_2^R, to playing an unrevealed row of higher maximum reward, a_1^R, for the last time in round t, then

it must have done so in its previous round, leading to a contradiction. In model $\mathbf{M_2}$, the human does not observe the payoffs of the row played by the robot before they plays their own action. Therefore, we can assume w.l.o.g. that when an unrevealed row is played, its reward is no larger than the maximum reward of that row, e.g., $C_1 \leqslant R_1$ if row 1 is played. Hence, if the optimal policy changes its action from a_2^{R} to a_1^{R} in round t when $x_t = (0, 1, \ldots)$, the inequality equivalent to eq. (5.7) must be

$$\begin{aligned} &C_1 + \alpha R_1 \cdot (T - t - 1) + (1 - \alpha) R_2 \cdot (T - t - 1) \\ &\quad > R_2 + R_2 \cdot (T - t - 1). \end{aligned} \tag{5.13}$$

After adding the slack variable, we get,

$$\begin{aligned} \rho_1 &\triangleq C_1 + \alpha R_1 \cdot (T - t - 1) + (1 - \alpha) R_2 \cdot (T - t - 1) \\ &= R_2 + R_2 \cdot (T - t - 1) + \epsilon \\ \implies C_1 + \alpha R_1 \cdot (T - t - 1) &= R_2 + \alpha R_2 \cdot (T - t - 1) + \epsilon. \end{aligned}$$

In *Case 1*, the expected payoff of playing a_1^{R} is at least: $C_1 + \alpha R_1 \cdot (T - t) + (1 - \alpha) U_t(\pi^{\mathrm{R}^*} | (0, 0, \ldots))$. The expected payoff of playing a_2^{R} is: $C_2 + \alpha \rho_1 + (1 - \alpha) U_t(\pi^{\mathrm{R}^*} | (0, 0, \ldots))$. We show that the first expression is larger than the second, i.e.,

$$\begin{aligned} &C_1 + \alpha R_1 \cdot (T - t) > C_2 + \alpha \rho_1 \\ \iff &\alpha R_1 + R_2 + \alpha R_2 \cdot (T - t - 1) + \epsilon \\ &\qquad > C_2 + \alpha R_2 + \alpha R_2 \cdot (T - t - 1) + \alpha \epsilon \\ \iff &\alpha R_1 + R_2 + \epsilon > C_2 + \alpha R_2 + \alpha \epsilon. \end{aligned}$$

The final inequality holds since $R_1 > R_2 \geqslant C_2$ and $0 < \alpha < 1$.

Similarly for *Case 2*, the expected payoff of playing a_1^{R} is at least:

$$\begin{aligned} &C_1 + \alpha R_1 \cdot (T - t) + (1 - \alpha) U_t(\pi^* | (0, 1, \ldots)) \\ &\quad \geqslant C_1 + \alpha R_1 \cdot (T - t) + (1 - \alpha) R_2 \cdot (T - t). \end{aligned}$$

On the other hand, the expected payoff of playing a_2^{R} is $R_2 + \rho_1$. We again show that the RHS of the first expression is larger than the second, i.e.,

$$\begin{aligned} &C_1 + \alpha R_1 \cdot (T - t) + (1 - \alpha) R_2 \cdot (T - t) > R_2 + \rho_1 \\ \iff &C_1 + \alpha R_1 \cdot (T - t - 1) + \alpha R_1 + (1 - \alpha) R_2 \cdot (T - t - 1) \\ &\qquad + (1 - \alpha) R_2 > R_2 + R_2 + R_2 \cdot (T - t - 1) + \epsilon \\ \iff &R_2 + R_2 \cdot (T - t - 1) + \epsilon + \alpha R_1 + (1 - \alpha) R_2 \\ &\qquad > R_2 + R_2 + R_2 \cdot (T - t - 1) + \epsilon \\ \iff &\alpha R_1 > \alpha R_2, \end{aligned}$$

which holds since $R_1 > R_2$ and $0 < \alpha < 1$. □

Algorithm 1 Optimal Policy: Full Observability

Input: matrix R, time horizon T, parameter α
Output: optimal action a_t^* in each round t
$U_t(x_t), a_t^*(x_t) =$ `OptPolicy`(x_t, t)
procedure `OptPolicy`(x_t, t)
  if $t > T$ **then**
    return `(0, None)`
  else
    if x_t has at least one 1 **then**
      find a row k^* s.t. $k^* \in \arg\max_{k:x_{t,k}=1} \max\ r_k$
      return $(\max r_{k^*} \times (T-t), k^*)$
    else
      find a row
      $i^* \in \arg\max_{k\in[m]} [\alpha(R_k + U_{t+1}(\mathbf{e}_k)) + (1-\alpha)(C_k + U_{t+1}(\mathbf{0}))]$
      and its value u_{i^*} (*for model* $\mathbf{M_1}$)
      OR
      find a row
      $i^* \in \arg\max_{k\in[m]} [C_k + \alpha U_{t+1}(\mathbf{e}_k) + (1-\alpha)U_{t+1}(\mathbf{0})]$
      and its value u_{i^*} (*for model* $\mathbf{M_2}$)
      return (u_{i^*}, i^*)
    end if
  end if
end procedure

5.2.4.3 Design of an efficient algorithm

As advertised, using Lemmas 1 and 2, we can easily prove the following theorem.

Theorem 1. *In models* $\mathbf{M_1}$ *and* $\mathbf{M_2}$*, an optimal policy can be computed in polynomial time.*

Indeed, the algorithm is specified as Algorithm 1. Here $\mathbf{e}_k$ denotes the m-dimensional standard unit vector in direction k. This algorithm runs in time polynomial in m and T since the inner `else` condition does not branch into two independent computations. This is because when at least one coordinate of x_t is 1, the inner `if` condition is met and the expected payoff in that case is computed without recursion. Therefore, in every round the number of computations is $O(m)$, and the algorithm has complexity $O(mT)$.

Figure 5.5: User performs a repeated table-clearing task with the robot. The robot fails intentionally in the beginning of the task, in order to reveal its capabilities to the human teammate.

5.2.5 *From Theory to Users*

We conduct a human subject experiment to evaluate the proposed model in a table-clearing task (fig. 5.5). We focus on the case where the human *learns from experience* (Models $\mathbf{M_2}, \mathbf{M_3}$). We are interested in showing that the policies computed using the partial adaptation model will result in better performance than policies that model the human as learning the best-response to all robot actions, rather than to only the robot action played.

5.2.6 *Manipulated Variables*

Observability. We used two settings — one where the robot does not directly observe whether the human has learned (section 5.2.3), and one where the robot observes directly whether the human has learned (section 5.2.4.2).
Adaptation. We compared the proposed partial adaptation model (fig. 5.6) with a model of complete adaptation, where the robot models the human as learning all rows of the payoff matrix with probability α after a row is played, instead of learning only the row played (fig. 5.7).

5.2.6.1 Hypothesis

We hypothesize that the robot policies that model the human as partially adapting to the robot will perform better than the policies that assume complete adaptation of the human to the robot.

5.2.6.2 Experiment Setting

Table-clearing task. We test the hypothesis in the table-clearing task of fig. 5.5, where a human and a robot collaborate to clear the table from objects. In this task, the human can take any of the following actions: {pick up any of the blue cups and place them on the blue bin, change the location of any of the bins, empty any of the bottles of water}. The robot can either remain idle or pick up any of the bottles from the table and move them to the red bin. The goal is to maximize the number of objects placed in the bins.

The human does not have in advance the following information about the robot: (1) the robot does not know the location of the green bin. Therefore, when the robot attempts to grab one of the bottles, it may push the green bin, dropping the blue bin off the table. (2) The robot will fail if it picks up the bottle that is farthest away from it, if that bottle has water in it. This is because of its motor torque limits.
Model parameters. This information is represented in the form of a payoff matrix R. The entries correspond to the number of objects

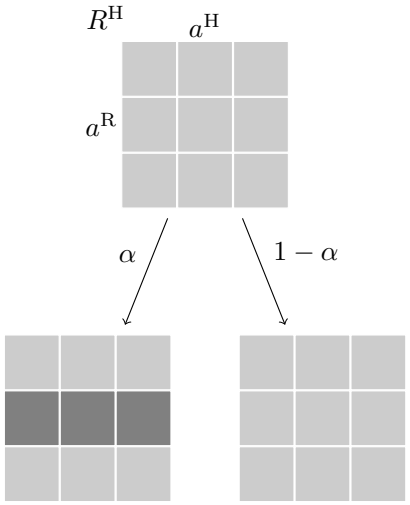


Figure 5.6: The robot reveals the row played (in this example row 2) with probability α.

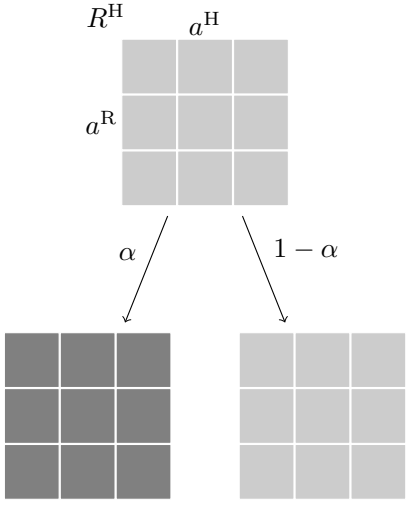


Figure 5.7: The robot reward matrix R is in dark shade and the human reward matrix R^{H} in light shade. The robot reveals its whole reward matrix with probability α.

in the bins after each human and robot action. Table 5.1 shows part of R; it includes only the subset of human actions that affect the outcome. For instance, if the robot starts moving towards the bottle that is closest to it (action 'Pick up closest') and the human does not move the green or blue bin out of the way, the robot will drop the blue bin off the table, together with any blue cups that the human has placed. Therefore, at the end of the task only the bottle will be cleared from the table, resulting in a reward of 1. If the robot attempts to pick up both bottles (action "pick up both") and the human does not empty the bottle of water before the robot grasps it, the robot will fail, resulting in a reward of 0. If the human has emptied the bottle and moved the blue bin (action "Clear cups & move bin & empty bottle"), the robot will successfully clear both bottles without dropping the bin, resulting in a reward of 4 (2 bottles in the red bin and 2 cups in the blue bin).

	Clear cups	Clear cups & move bin	Clear cups & move bin & empty bottle
Noop	2	2	2
Pick up closest	1	3	3
Pick up both	0	0	4

Table 5.1: Part of payoff matrix R for table-clearing task. The table includes only the subset of human actions that affect performance.

In the beginning of the task, we assume that the human response to all robot actions will be "Clear cups"; since the human has not observed the robot dropping the bin or failing to pick up the bottle, they have no reason to move the bin or empty the bottle of water. We also assume that they do not learn any payoffs if the robot remains idle ("Noop" action). We set the probability of learning $\alpha = 0.9$, since we expected most participants to learn the best-response to the robot actions after observing the outcome of their actions.

Procedure. The experimenter first explained the task to the participants and informed them about the actions that they could take, as well as about the robot actions. Participants were told that the goal was to maximize the number of objects placed in the bins at each round. They performed the task three times ($T = 3$). In the full observability setting, the experimenter asked the participants after each round, what would their action be if the robot did the same action in the next round. The experimenter then inputted their response (learned or not learned) into the program that executed the policy. When the robot failed to pick up the bottle, the experimenter informed them that the robot had failed. Participants were told that the error message displayed in the terminal was: "The torque of the robot motors exceeded their limits." This is the generic output of our

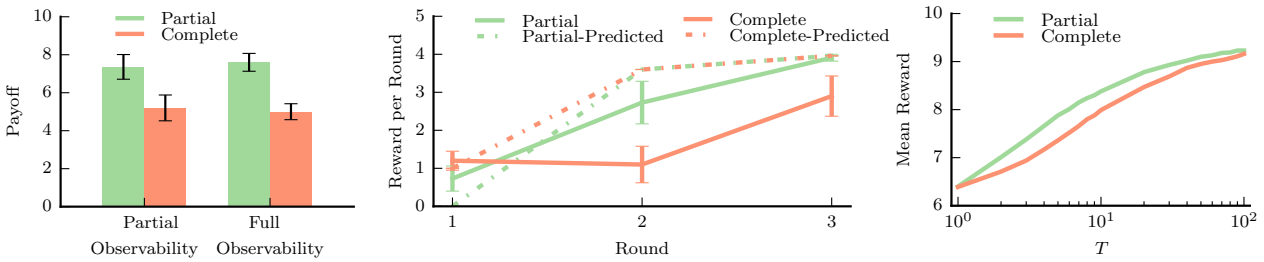


Figure 5.8: (left) Accumulated reward over 3 trials of the table-clearing task for all four conditions. (center) Predicted and actual reward by the partial and complete adaptation policies in the partial observability setting. (right) Mean reward over time horizon T for simulated runs of the complete and partial adaptation policies in the partial observability setting. The gain in performance from the partial adaptation model decreases for large values of T. The x-axis is in logarithmic scale.

ROS-based hardware interface, when the measured torques exceed the manufacturer limits. We added a short, general explanation about how torque is related to distance and applied force. At the end, participants answered open-ended questions about their experience in the form of a video-taped interview.

5.2.7 *Subject Allocation*

We recruited 60 participants from a university campus. We chose a between-subjects design in order to avoid biasing users towards policies from previous conditions.

5.2.8 *Results and Discussion*

Analysis. We evaluate team performance by the accumulated reward over the three rounds of the task (fig. 5.8-left). We observe that the mean reward in the partial adaptation policy was 42% higher than that of the complete adaptation policy in the partial observability setting, and 52% higher than that of the complete adaptation policy in the full observability setting. A factorial ANOVA showed no significant interaction effects between the observability and adaptation factors. The test showed a statistically significant main effect of adaptation ($F(1,56) = 18.58, p < 0.001$), and no significant main effect of observability. These results support our hypothesis.

The difference in performance occurred because in the complete adaptation model the robot erroneously assumed that the human had learned the best-response to the "Pick up both" action, after the robot played the row "Pick up closest." In this section, we examine the partial and complete adaptation policies in the *partial-observability* setting. The interpretation of the robot actions in the *full-observability* setting is similar. The robot chooses the action "Pick up both" for round $T = 1$ (as well as for $T = 2, 3$) in the partial adaptation condition [8], since the loss of receiving zero reward at $T = 1$ is outweighed by the rewards of 4 in subsequent rounds, if the human learns the best-response to that action, which occurs with high probability

[8] Unless specified otherwise, for the rest of this section we refer to the partial observability level of the observability factor.

($\alpha = 0.9$). On the other hand, the robot in the complete adaptation condition takes the action "Pick up closest" at $T = 1$ and "Pick up both" at $T = 2$ and $T = 3$. This is because the model assumes that the human will learn the best-response for all robot actions if the robot plays either "Pick up closest" or "Pick up both", and the predicted reward of 1 for the action "Pick up closest" is higher than the reward of 0 for "Pick up both" at $T = 1$.

Fig. 5.8-center shows the expected immediate reward predicted by the partial and complete adaptation model for each round in the partial observability setting, and the actual reward that participants received. We see that the immediate reward in the complete adaptation condition at $T = 2$ was significantly lower than the predicted one. The reason is that six participants out of 10 in that condition did not infer at $T = 1$ that the robot was unable to pick up the second bottle and did not empty the bottle at $T = 2$, which was the best-response action. From the four participants that emptied the bottle, two of them justified their action by stating that "there was enough time to empty the bottle" before the robot would grab it. The same justification was given by three participants out of eleven in the partial adaptation condition, who emptied the bottle at $T = 1$ without knowing that this was required for the robot to be able to pick it up. This caused the actual reward to be higher than its predicted value of 0. Additionally, the actual reward at $T = 2$ was lower than the predicted value. We attribute this to the fact that 73% of participants learned the best-response for the robot action (emptying the bottle that was farthest away) in that round, whereas the predicted value assumed a probability of learning $\alpha = 0.9$. In $T = 3$, the actual reward matched the prediction closely, since all participants eventually learned that they should empty the bottle.

Generalizability of the results. The results discussed above are compelling in that they arise from an actual human-subject experiment, but they are limited to one task. We are interested in showing — via simulations — that the proposed model performs well for a variety of tasks. We randomly generated instances of the reward matrix R and α values and simulated runs of the partial and complete adaptation policies for increasing time horizons T. The simulated humans partially adapted to the robot, and the robot did not observe whether they learned. For each value of T, we randomly sampled 1000 reward matrices R and simulated 100 policy runs for each sampled instance of R. Fig. 5.8-right shows the reward averaged over the number of rounds T, policy runs and instances of R. For $T = 1$, the mean reward is the same for both models, since there is no adaptation. The partial adaptation policies consistently outperform the complete adaptation ones for a large range of T. For large values of T the per-

formance difference decreases. This is because the human eventually learns the true payoffs and the gain from playing the true best response outweighs the initial loss caused by the complete adaptation model.

Selection of α. We note that the α value, which represents the probability that the human learns the true robot capabilities, is task and population-dependent. In our experiment, participants were recruited from a university campus, and most of them were able to infer that they should empty the bottle, after observing the robot failing and being notified that "the robot exceeded its torque limits." Different participant groups may require a different α value. The value of α could also vary for different robot actions; we conjecture that our theoretical results hold also when there is a different adaptation probability α_i for each row i of the payoff matrix, which we leave as future work.

5.2.9 Conclusion

We presented a game-theoretic model of human partial adaptation to the robot. The robot used this model to decide optimally between taking actions that reveal its capabilities to the human and taking the best action given the information that the human currently has. We proved that under certain observability assumptions, the optimal policy can be computed efficiently. Through a human subject experiment, we demonstrated that policies computed with the proposed model significantly improved human-robot team performance, compared to policies that assume complete adaptation of the human to the robot.

While our model assumes that the human may learn only the entries of the row played by the robot, there are cases where a robot action may affect entries that are associated with other actions, as well. For instance, Cha et al. [2015] have shown that conversational speech can affect human perception of robot capability in physical tasks. Future work includes exploring the structure of probabilistic graphical models of human adaptation, and using the theoretical insights from this work to develop efficient algorithms for the robot.

5.3 Discussion

This chapter described two models of human adaptation, where the human changes its behavior based on the robot's actions. The robot leverages this to communicate information to its human teammate and guide them towards better ways of doing the task. In these models, we have assumed the human state $\theta \in \Theta$ to be *fully observable*. In

the next chapter, we relax this assumption and show that the robot can infer the unknown human state of a new human teammate and build a model of human adaptation online, through interaction.

6
Mutual Adaptation

In our models of human adaptation, we have assumed that the robot knows the type θ of the human, which parameterizes the human policy π^H and the human reward function R^H. However, our studies have shown that there is a large variability among different types θ. Additionally, the type θ of a new human worker is typically unknown in advance to the robot and it cannot be fully observed. In this chapter [1] we relax the assumption of a known θ for the human. Instead, we treat θ as a latent variable in a partially observable stochastic process, in particularly a mixed-observability Markov decision process, which has been shown to achieve significant computational efficiency [Ong et al., 2010]. This allows the robot to take information seeking actions to infer online the parameter θ, which specifies how the human policy π^H is affected by the robot's own actions. As a result, human and robot *mutually adapt* to each other; the robot builds online a model of how the human adapts to the robot by inferring their type θ, and adapts its own actions in return.

[1] Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddhartha Srinivasa. Formalizing human-robot mutual adaptation: A bounded memory model. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016; Stefanos Nikolaidis, Yu Xiang Zhu, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in shared autonomy. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017c; and Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research (IJRR)*, 2017a

In section 5.1 of chapter 5, we instantiated the type θ of the human as the human mode m^H and the human adaptability α, so that $\theta = (m^H, \alpha)$. In the mutual adaptation formalism of section 6.1, we keep the full observability assumption for the mode m^H, and treat the human adaptability as a latent variable. In section 6.2, we relax the full observability assumption for the mode m^H.

In this chapter, we assume that the human adaptability is constant throughout the task. We break this assumption in the next chapter.

6.1 Collaboration

Work done in collaboration with David Hsu and Anton Kuznetsov.

We use as application the table-carrying task of chapter 5 (fig. 6.1). We model the human policy using the Bounded memory Adaptation model of section 5.1, chapter 5, and treat the human adaptability as a latent variable in a partially observable stochastic process. This enables the robot to infer the human adaptability online through

interaction, and adapt its own actions in return. Fig. 6.2 shows examples of human and robot behaviors for three simulated humans in the task. The robot estimates the unknown human adaptability through interaction. User 1 is fully non-adaptable with $\alpha = 0$. The robot infers this after two steps of interaction and switches its action to comply with the human preference. User 3 is fully adaptable with $\alpha = 1$ and switches to accommodate the robot preference after one step of interaction. User 2 is adaptable with $\alpha = 0.75$. After several steps, the robot gets a good estimate on the human adaptability level and guides the human to the preferred strategy.

We want to emphasize here that the robot executes a single policy that adapts to different human behaviors. If the human is non-adaptable, the robot complies to the human's preferred strategy. Otherwise, the robot guides the human towards a better strategy.

We are interested in studying whether a robot, under our proposed approach, is able to guide human partners towards a better collaboration strategy, sometimes against their initial preference, while still retaining their trust. We conducted a human subject experiment online via video playback ($n = 69$) on the simulated table carrying task (fig. 6.1). In the experiment, participants were significantly more likely to adapt, when working with the robot utilizing our mutually adaptive approach, compared with the robot that cross-trained with the participants. Additionally, the participants found that the mutually adaptive robot has performance not worse than the cross-trained robot. Finally, the participants found that the mutually adaptive robot was more trustworthy than the robot executing a fixed strategy optimal in task performance, but ignoring the human preference.

We are also interested in how adaptability and trust change over time. We hypothesized that trust in the mutually adaptive robot increases over time for non-adaptable participants, as previous work suggests that robot adaptation significantly improves perceived robot trustworthiness [Shah et al., 2011], and that the increase in trust results in subsequent increased likelihood of human adaptation to the robot. A human subject experiment on repeated table-carrying tasks ($n = 43$) did not support this hypothesis.

To study the generality of our model, we hypothesized that non-adaptable participants in the table-carrying task would be less likely to adapt in a different collaborative task. A follow-up human subject experiment with a hallway-crossing task ($n = 58$) confirmed the hypothesis.

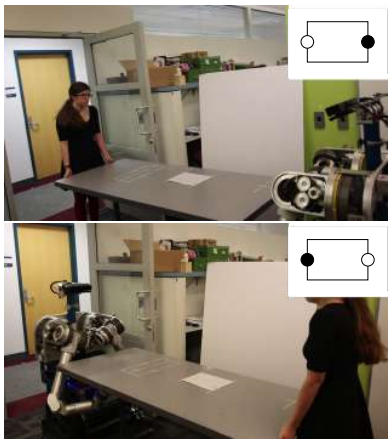


Figure 6.1: A human and a robot collaborate to carry a table through a door. (top) The robot prefers facing the door (Goal A), as it has a full view of the door. (bottom) The robot faces away from the door (Goal B).

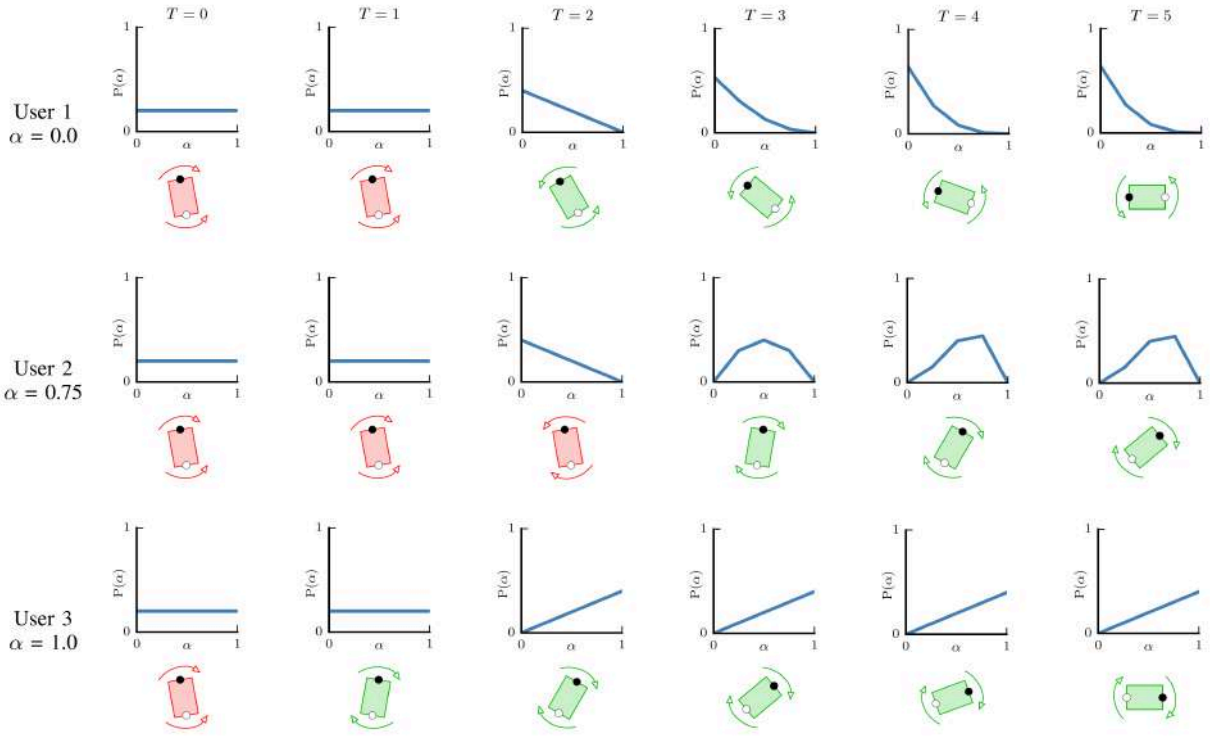


Figure 6.2: Sample runs on the human-robot table-carrying task, with three simulated humans of adaptability level α=0, 0.75, and 1. A fully adaptable human has α=1, while a fully non-adaptable human has α=0. In each case, the upper row shows the probabilistic estimate on α over time. The lower row shows the robot and human actions over time. Red color indicates human (white dot) and robot (black dot) disagreement in their actions, in which case the table does not move. The columns indicate successive time steps. User 1 is non-adaptable, and the robot complies with his preference. User 2 and 3 are adaptable to different extent. The robot successfully guides them towards a better strategy.

6.1.1 *Robot Planning*

In this section we describe the integration of BAM in the robot decision making process using a MOMDP formulation. A MOMDP uses proper factorization of the observable and unobservable state variables $S : X \times Y$ with transition functions $\mathcal{T}_x$ and $\mathcal{T}_y$, reducing the computational load [Ong et al., 2010]. The set of observable state variables is $X : X^{\mathrm{w}} \times M^k \times M^k$, where X^{w} is the finite set of task-steps that signify progress towards task completion and M is the set of modal policies followed by the human and the robot in a history length k. The partially observable variable y is identical to the human adaptability α. We assume finite sets of human and robot actions A^{H} and A^{R}, and we denote as π^{H} the stochastic human policy. The latter gives the probability of a human action a^{H} at state s, based on the BAM human adaptation model.

The belief update becomes:

$$\begin{aligned} b'(y') =& \eta O(s', a^{\mathrm{R}}, o) \sum_{y \in Y} \sum_{a^{\mathrm{H}} \in A^{\mathrm{H}}} \mathcal{T}_x(s, a^{\mathrm{R}}, a^{\mathrm{H}}, x') \\ & \mathcal{T}_y(s, a^{\mathrm{R}}, a^{\mathrm{H}}, s') \pi^{\mathrm{H}}(x, y, a^{\mathrm{H}}) b(y) \end{aligned} \tag{6.1}$$

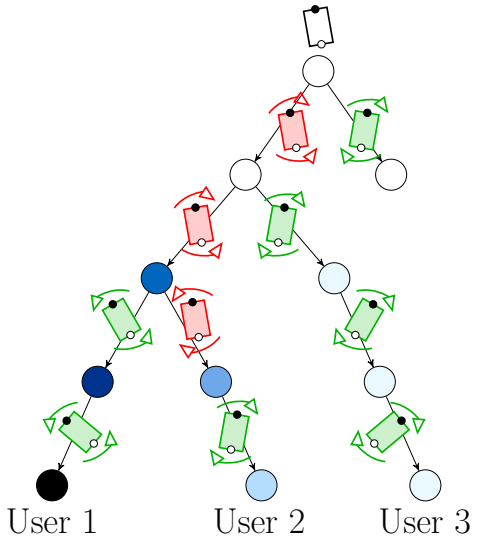


Figure 6.3: Different paths on MOMDP policy tree for human-robot (white/black dot) table-carrying task. The circle color represents the belief on α, with darker shades indicating higher probability for smaller values (less adaptability). The white circles denote a uniform distribution over α. User 1 is inferred as non-adaptable, whereas Users 2 and 3 are adaptable.

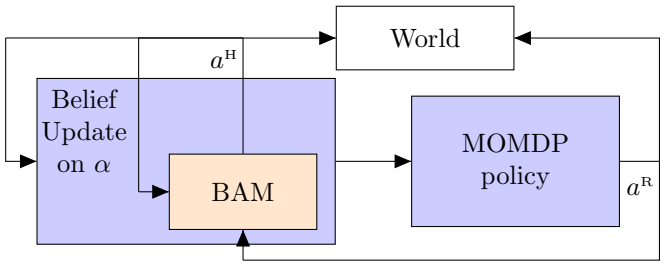


Figure 6.4: Integration of BAM into MOMDP formulation.

We use a point-based approximation algorithm to solve the MOMDP for a robot policy π^R that takes into account the robot belief on the human adaptability, while maximizing the agent's expected total reward.

The policy execution is performed online in real time and consists of two steps (fig. 6.4). First, the robot uses the current belief to select the action a^R specified by the policy. Second, it uses the human action a^H to update the belief on α (eq. 6.1). Fig. 6.3 presents the paths on the MOMDP policy tree that correspond to the simulated user behaviors presented in fig. 6.2.

6.1.2 *Human Subject Experiment*

We revisit the table-carrying task of chapter 5. We are interested in showing that integrating BAM into the robot decision making can lead to more efficient policies than state-of-the-art human-robot team training practices, while maintaining human satisfaction and trust.

On one extreme, we can "fix" the robot policy so that the robot always moves towards the optimal —with respect to some objective performance metric —goal, ignoring human adaptability. This will

force all users to adapt, since this is the only way to complete the task. However, we hypothesize that this will significantly impact human satisfaction and trust in the robot. On the other extreme, we can efficiently learn the human preference [Nikolaidis and Shah, 2013]. This can lead to the human-robot team following a sub-optimal policy, if the human has an inaccurate model of the robot capabilities. We have, therefore, two control conditions: one where participants interact with the robot executing a fixed policy, always acting towards the optimal goal, and one where the robot learns the human preference. We show that the proposed formalism achieves a trade-off between the two: When the human is non-adaptable, the robot follows the human strategy. Otherwise, the robot insists on the optimal way of completing the task, leading to significantly better policies compared to learning the human preference.

6.1.2.1 Independent Variables

We had three experimental conditions, which we refer to as "Fixed," "Mutual-adaptation" and "Cross-training."

Fixed session The robot executes a fixed policy, always acting towards the optimal goal. In the table-carrying scenario, the robot keeps rotating the table in the clockwise direction towards Goal A, which we assume to be optimal (fig. 6.1). The only way to finish the task is for the human to rotate the table in the same direction as the robot, until it is brought to the horizontal configuration of Goal A.

Mutual-adaptation session The robot executes the MOMDP policy computed using the proposed formalism. The robot starts by rotating the table towards the optimal goal (Goal A). Therefore, adapting to the robot strategy corresponds to rotating the table to the optimal configuration.

Cross-training session Human and robot train together using the human-robot cross-training algorithm [Nikolaidis and Shah, 2013]. The algorithm consists of a forward phase and a rotation phase. In the forward phase, the robot executes an initial policy, which we choose to be the one that leads to the optimal goal. Therefore, in the table-carrying scenario, the robot rotates the table in the clockwise direction towards Goal A. In the rotation phase, human and robot switch roles, and the human inputs are used to update the robot reward function. After the two phases, the robot policy is recomputed.

6.1.2.2 Hypotheses

H1 *Participants will agree more strongly that HERB is trustworthy, and will be more satisfied with the team performance in the Mutual-adaptation condition, compared to working with the robot in the Fixed condition.* We

expected users to trust more the robot with the learned MOMDP policy, compared with the robot that executes a fixed strategy ignoring the user's willingness to adapt. In prior work, a task-level executive that adapted to the human partner significantly improved perceived robot trustworthiness [Shah et al., 2011]. Additionally, working with a human-aware robot that adapted its motions had a significant impact on human satisfaction [Lasota and Shah, 2015].

H2 *Participants are more likely to adapt to the robot strategy towards the optimal goal in the Mutual-adaptation condition, compared to working with the robot in the Cross-training condition*. The computed MOMDP policy enables the robot to infer online the adaptability of the human and guides adaptable users towards more effective strategies. Therefore, we posited that more subjects would change their strategy when working with the robot in the Mutual-adaptation condition, compared with cross-training with the robot. We note that in the Fixed condition all participants ended up changing to the robot strategy, as this was the only way to complete the task.

H3 *The robot performance as a teammate, as perceived by the participants in the Mutual-adaptation condition, will not be worse than in the Cross-training condition.* The learned MOMDP policy enables the robot to follow the preference of participants that are less adaptable, while guiding towards the optimal goal participants that are willing to change their strategy. Therefore, we posited that this behavior would result on a perceived robot performance not inferior to that achieved in the Cross-training condition.

6.1.2.3 Experiment Setting: A Table-Carrying Task

We first instructed participants in the task and asked them to choose one of the two goal configurations (fig. 6.1), as their preferred way of accomplishing the task. To prompt users to prefer the sub-optimal goal, we informed them about the starting state of the task, where the table was slightly rotated in the counter-clockwise direction, making the sub-optimal Goal B appear closer. Once the task started, the user chose the rotation actions by clicking on buttons on a user interface (fig. 6.5). If the robot executed the same action, a video played showing the table rotation. Otherwise, the table did not move and a message appeared on the screen notifying the user that they tried to rotate the table in a different direction than the robot. In the Mutual-adaptation and Fixed conditions participants executed the task twice. Each trial ended when the team reached one of the two goal configurations. In the Cross-training condition, participants executed the forward phase of the algorithm in the first trial and the rotation phase, where human and robot switched roles, in the second

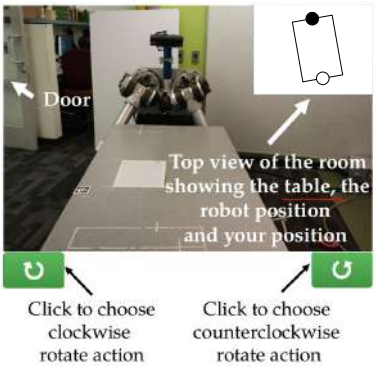


Figure 6.5: UI with instructions

trial. We found that in this task one rotation phase was enough for users to successfully demonstrate their preference to the robot. Following [Nikolaidis and Shah, 2013], the robot executed the updated policy with the participant in a task-execution phase that succeeded the rotation phase.

We asked all participants to answer a post-experimental questionnaire that used a five-point Likert scale to assess their responses to working with the robot (Table 6.1). We used the composite measures proposed by [Hoffman, 2013]. Questions 1 and 3 are from Hoffman's measure of "Robot Teammate Traits," while questions 4-6 are from Hoffman's adaptation of the "Working Alliance Index" for human-robot teams. Items 7-8 were proposed by [Gombolay et al., 2014] as additional metrics of team-fluency. We added questions 9-10 based on our intuition. Participants also responded to open-ended questions about their experience.

Table 6.1: Post-experimental questionnaire.

Q1:	"HERB is trustworthy."
Q2:	"I trusted HERB to do the right thing at the right time."
Q3:	"HERB is intelligent."
Q4:	"HERB perceived accurately what my goals are."
Q5:	"HERB did not understand how I wanted to do the task."
Q6:	"HERB and I worked towards mutually agreed upon goals."
Q7:	"I was satisfied with HERB and my performance."
Q8:	"HERB and I collaborated well together."
Q9:	"HERB made me change my mind during the task."
Q10:	"HERB's actions were reasonable."

6.1.2.4 Subject Allocation

We chose a between-subjects design in order to not bias the users with policies from previous conditions. We recruited participants through Amazon's Mechanical Turk service, all from the United States, aged 18-65 and with approval rates higher than 95%. Each participant was compensated $0.50. Since we are interested in exploring human-robot mutual adaptation, we disregarded participants that had as initial preference the robot goal. To ensure reliability of the results, we asked all participants a control question that tested

their attention to the task and eliminated data associated with wrong answers to this question, as well as incomplete data. To test their attention to the Likert questionnaire, we included a negative statement with the opposite meaning to its positive counterpart and eliminated data associated with positive or negative ratings to both statements, resulting in a total of 69 samples.

6.1.2.5 MOMDP Model

The observable state variables x of the MOMDP formulation were the discretized table orientation and the human and robot modes for each of the three previous time-steps. We specified two modal policies, each deterministically selecting rotation actions towards each goal. The size of the observable state-space X was 734 states. We set a history length $k = 3$ in BAM. We additionally assumed a discrete set of values of the adaptability $\alpha : \{0.0, 0.25, 0.5, 0.75, 1.0\}$. Although a higher resolution in the discretization of α is possible, we empirically verified that 5 values were enough to capture the different adaptive behaviors observed in this task. The total size of the MOMDP state-space was $5 \times 734 = 3670$ states. The human and robot actions a^H, a^R were deterministic discrete table rotations. We set the reward function R to be positive at the two goal configurations based on their relative cost, and 0 elsewhere. We computed the robot policy using the SARSOP solver [Kurniawati et al., 2008], a point-based approximation algorithm which, combined with the MOMDP formulation, can scale up to hundreds of thousands of states [Bandyopadhyay et al., 2013].

6.1.3 Results and Discussion

6.1.3.1 Subjective Measures

We consider hypothesis **H1**, that participants will agree more strongly that HERB is trustworthy, and will be more satisfied with the team performance in the Mutual-adaptation condition, compared to working with the robot in the Fixed condition. A two-tailed Mann-Whitney-Wilcoxon test showed that participants indeed agreed more strongly that the robot utilizing the proposed formalism is trustworthy ($U = 180, p = 0.048$). No statistically significant differences were found for responses to statements eliciting human satisfaction: "I was satisfied with the robot and my performance" and "HERB and I collaborated well together." One possible explanation is that participants interacted with the robot through a user interface for a short period of time, therefore the impact of the interaction on user satisfaction was limited.

We were also interested in observing how the ratings in the first two conditions varied, depending on the participants' willingness to change their strategy. Therefore, we conducted a post-hoc experimental analysis of the data, grouping the participants based on their adaptability. Since the true adaptability of each participant is unknown, we estimated it by the mode of the belief formed by the robot at the end of the task on the adaptability α:

$$\hat{\alpha} = \arg\max_{\alpha} b(\alpha) \tag{6.2}$$

We considered only users whose mode was larger than a confidence threshold and grouped them as *very adaptable* if $\hat{\alpha} > 0.75$, *moderately adaptable* if $0.5 < \hat{\alpha} \leq 0.75$ and *non-adaptable* if $\hat{\alpha} \leq 0.5$. fig. 6.6 shows the participants' rating of their agreement on the robot trustworthiness, as a function of the participants' group for the two conditions. In the Fixed condition there was a trend towards positive correlation between the annotated robot trustworthiness and participants' inferred adaptability (Pearson's $r = 0.452$, $p = 0.091$), whereas there was no correlation between the two for participants in the Mutual-adaptation condition ($r = -0.066$). We attribute this to the MOMDP formulation allowing the robot to reason over its estimate on the adaptability of its teammate and change its own strategy when interacting with non-adaptable participants, therefore maintaining human trust.

In this work, we elicited trust at the end of the task using participants' rating of their agreement to the statement "HERB is trustworthy," which has been used in previous work in human-robot collaboration ([Shah et al., 2011, Hoffman, 2013]). We refer the reader to [Desai, 2012, Kaniarasu et al., 2013, Xu and Dudek, 2015, Yanco et al., 2016] for approaches on measuring trust in real-time.

We additionally coded participants' open-ended comments about their experience with working with HERB, and grouped them based on the content and the sentiment (positive, negative or neutral). Table 6.2 shows the different comments and associated sentiments, and fig. 6.7 illustrates the participants' ratio for each comment. We note that 20% of participants in the Fixed condition had a negative opinion about the robot behavior, noting that "[HERB] was poorly designed," and that probably "robot development had not been mastered by engineers" (C8 in Table 6.2). On the other hand, 26% of users in the Mutual-adaptation condition noted that the robot "attempted to anticipate my moves" and "understood which way I wanted to go" (C2). Several adaptable participants in both conditions commented that "[HERB] was programmed to move this way" (C5), while some of them attempted to justify HERB's actions, stating that it "was probably unable to move backwards" (C4).

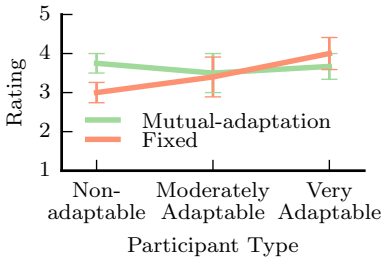


Figure 6.6: Rating of agreement to statement "HERB is trustworthy." Note that the figure does not include participants, whose mode of the belief on their adaptability was below a confidence threshold.

	Description	Sentiment
C1	"The robot followed my instructions."	Positive
C2	"The robot adapted to my actions."	Positive
C3	"The robot wanted to be efficient."	Positive
C4	"The robot was unable to move."	Neutral
C5	"The robot was programmed this way."	Neutral
C6	"The robot wanted to face the door."	Neutral
C7	"The robot was stubborn."	Negative
C8	"The robot was poorly programmed."	Negative

Table 6.2: Participants' comments and associated sentiments.

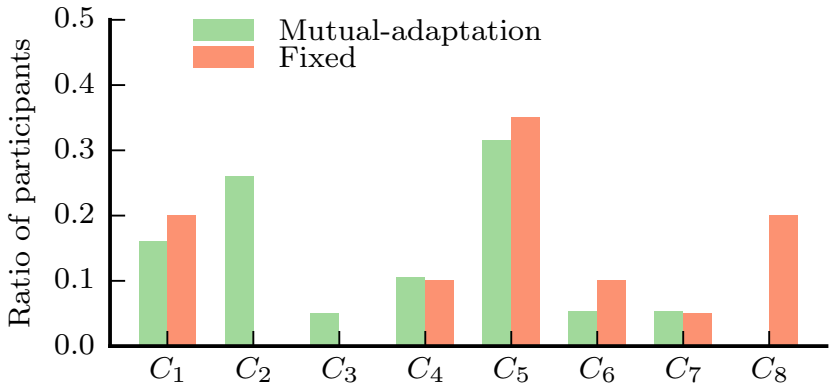


Figure 6.7: Ratio of participants per comment for the Mutual-adaptation and Fixed conditions.

Recall hypothesis **H3**: that the robot performance as a teammate in the Mutual-adaptation condition, as perceived by the participants, would not be worse than in the Cross-training condition. We define "not worse than" similarly to [Dragan et al., 2013] using the concept of "non-inferiority" [Lesaffre, 2008]. An one-tailed unpaired t-test for a non-inferiority margin $\Delta = 0.5$ and a level of statistical significance $\alpha = 0.025$ showed that participants in the Mutual-adaptation condition rated their satisfaction on robot performance ($p = 0.006$), robot intelligence ($p = 0.024$), robot trustworthiness ($p < 0.001$), quality of robot actions ($p < 0.001$) and quality of collaboration ($p = 0.002$) not worse than participants in the Cross-training condition. With Bonferroni corrections for multiple comparisons, robot trustworthiness, quality of robot actions and quality of collaboration remain significant. This supports hypothesis **H3** of section 6.1.2.2.

6.1.3.2 Quantitative Measures

To test hypothesis **H2**, we consider the ratio of participants that changed their strategy to the robot strategy towards the optimal goal

in the Mutual-adaptation and Cross-training conditions. A change was detected when the participant stated as preferred strategy a table rotation towards Goal B (fig. 6.1) , but completed the task in the configuration of Goal A in the final trial of the Mutual-adaptation session, or in the task-execution phase of the Cross-training session. As fig. 6.8 shows, 57% of participants adapted to the robot in the Mutual-adaptation condition, whereas 26% adapted to the robot in the Cross-training condition. A Pearson's chi-square test showed that the difference is statistically significant ($\chi^2(1, N = 46) = 4.39, p = 0.036$). Therefore, participants that interacted with the robot of the proposed formalism were more likely to switch to the robot strategy towards the optimal goal, than participants that cross-trained with the robot, which supports our hypothesis.

In section 6.1.3.3, we discuss the robot behavior for different values of history length k in BAM.

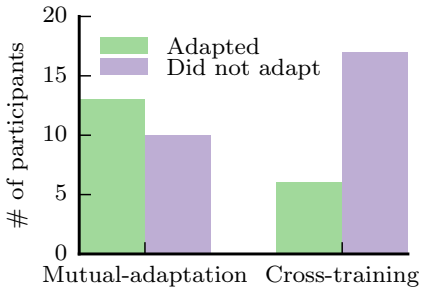


Figure 6.8: Number of participants that adapted to the robot for the Mutual-adaptation and Cross-training conditions.

6.1.3.3 Selection of History Length

The value of k in BAM indicates the number of time-steps in the past that we assume humans consider in their decision making on a particular task, ignoring all other history. Increasing k results in an exponential increase of the state space size, with large values reducing the robot responsiveness to changes in the human behavior. On the other hand, very small values result in unrealistic assumptions on the human decision making process.

To illustrate this, we set $k = 1$ and ran a pilot study of 30 participants through Amazon-Turk. Whereas most users rated highly their agreement to questions assessing their satisfaction and trust in the robot, some participants expressed their strong dissatisfaction with the robot behavior. This occurred when human and robot oscillated back and forth between modes, similarly to when two pedestrians on a narrow street face each other and switch sides simultaneously until they reach an agreement. In this case, which occurred in 23% of the samples, when the human switched back to their initial mode, which was also the robot mode of the previous time-step, the robot incorrectly inferred them as adaptable. However, the user in fact resumed their initial mode followed before two time-steps, implying a tendency for non-adaptation. This is a case where the 1-step bounded memory assumption did not hold.

In the human subject experiment of section 6.1.2, we used $k = 3$, since we found this to describe accurately the human behavior in this task. Fig. 6.9 shows the belief update and robot behavior for $k = 1$ and $k = 3$, in the case of mode oscillation. At $T = 1$, after the first disagreement and in the absence of any previous history, the

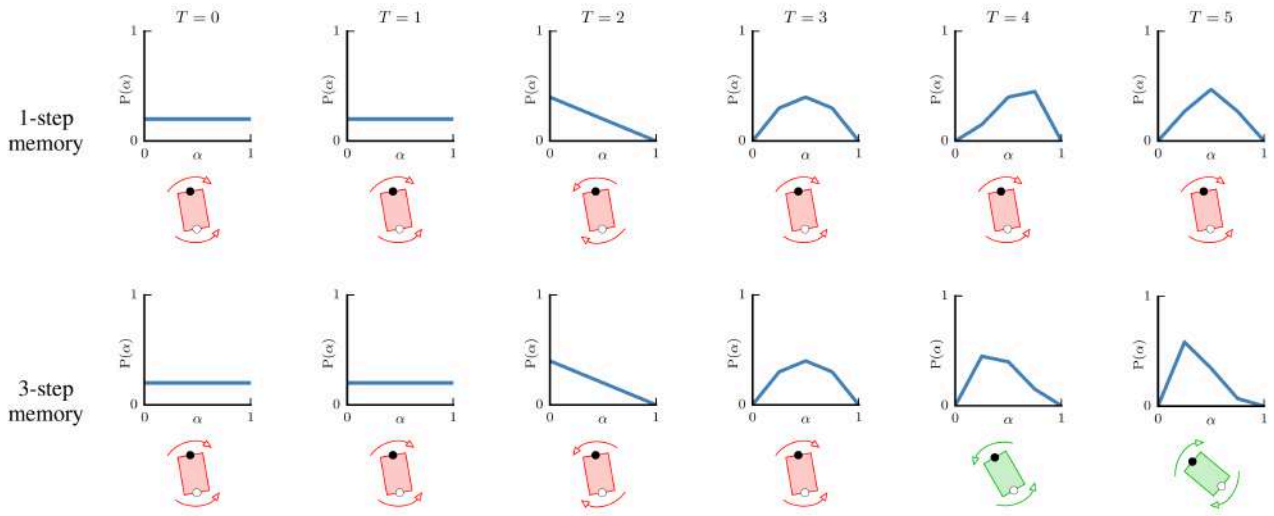


Figure 6.9: Belief update and table configurations for the 1-step (top) and 3-step (bottom) bounded memory models at successive time-steps.

belief remains uniform over α. The human (white dot) follows their modal policy from the previous time-step, therefore at $T = 2$ the belief becomes higher for smaller values of α in both models (lower adaptability). At $T = 2$, The robot (black dot) adapts to the human and executes the human modal policy. At the same time, the human switches to the robot mode, therefore at $T = 3$ the probability mass moves to the right. At $T = 3$, the human switches back to their initial mode. In the 3-step model the resulting distribution at $T = 4$ has a positive skewness: the robot estimates the human to be non-adaptable. In the 1-step model the robot incorrectly infers that the human adapted to the robot mode of the previous time-step, and the probability distribution has a negative skewness. At $T = 4, 5$, the robot in the 3-step trial switches to the human modal policy, whereas in the 1-step trial it does not adapt to the human, who insists on their mode.

6.1.3.4 Discussion

This online study in the table-carrying task seems to suggest that the proposed formalism enables a human-robot team to achieve more effective policies, compared to state-of-the-art human-robot team training practices, while achieving subjective ratings on robot performance and trust that are comparable to those achieved by these practices. It is important to note that the comparison with the human-robot cross-training algorithm is done in the context of human adaptation. Previous work [Nikolaidis and Shah, 2013] has shown that switching roles can result in significant benefits in

team fluency metrics, such as human idle time and concurrent motion [Hoffman and Breazeal, 2007], when a human executes the task with an actual robot. Additionally, the proposed formalism assumes as input a set of modal policies, as well as a quality measure associated with each policy. On the other hand, cross-training requires only an initialization of a reward function of the state space, which is then updated in the rotation phase through interaction. It would be very interesting to explore a hybrid approach between learning the reward function and guiding the human towards an optimal policy, but we leave this for future work.

6.1.3.5 Information-Seeking Behavior

We observe that in the experiments, the robot always starts moving towards the optimal goal, until it is confident that the human is non-adaptable, in which case it adapts to the human. The MOMDP chooses whether the robot should adapt or not, based on the estimate of the human adaptability, the rewards of the optimal and suboptimal goal and the discount factor.

In the general case, information-seeking actions can occur at any point during the task. For instance, in a multi-staged task, where information gathering costs differently in different stages (i.e. moving a table out of the room / through a narrow corridor), the robot might choose to disagree with the human in a stage where information-seeking actions are cheap, even if the human follows an optimal path in that stage.

6.1.3.6 Generalization to Complex Tasks

The presented table-carrying task can be generalized without significant modifications in the proposed mathematical model, with the cost of increasing the size of the state-space and action-space. In particular, we made the assumptions: (1) discrete time-steps, where human and robot apply torques causing a fixed table-rotation. (2) binary human-robot actions. (3) fully observable modal policies. We discuss how we can relax these assumptions:

1. We can approximate a continuous-time setting by increasing the resolution of the time discretization. Assuming a constant displacement per unit time v and a time-step dt, the size of the state-space increases linearly with $(1/dt)$: $O(|X^w||M|^{2k}) = O((\theta_{max} - \theta_{min}) * (1/v) * (1/dt) * |M|^{2k})$, where θ is the rotation angle of the table.

2. The proposed formalism is not limited to binary actions. For instance, we can allow torque inputs of different magnitudes. The

action-space of the MOMDP increases linearly with the number of possible inputs.

3. While we assumed that the modal policies are fully observable, an assumption that enables the human and the robot to infer a mode by observing an action, in the general case different modal policies may share the same action selection in some states, which would make them undeterminable. In this case, the proposed formalism can be generalized to include the human modal policy as additional latent variable in the MOMDP. Similarly, we can model the human as inferring a probability distribution over modes from the recent history, instead of inferring the robot mode with the maximum frequency count (eq. 5.2 in section 5.1.2). We leave this for future work.

Finally, we note that the presented formalism assumes that the world-state, representing the current task-step, is fully observable, and that human and robot have a known set of actions. This assumption holds for tasks with clearly defined objectives and distinct task-steps. In section 6.1.5, we apply our formalism in the case where human and robot cross a hallway and coordinate to avoid collision, and the robot guides the human towards one side of the corridor. Applicable scenarios include also a wide range of manufacturing tasks (e.g. assembly of airplane spars), where the goal and important concepts, such as tolerances and completion times, are defined in advance, but the sequencing of subtasks is flexible and can vary based on the individualized style of the mechanic [Nikolaidis et al., 2015b]. In these scenarios, the robot could lead the human to strategies that require less time or resources.

6.1.4 Adaptability in Repeated Trials

Previous work by Shah et al. [2011] has shown that robot adaptation significantly improves perceived robot trustworthiness. Therefore, we hypothesized that trust in the mutually adaptation condition would increase over time for non-adaptable participants, and that this increase in trust would result in a subsequent increased likelihood of human adaptation to the robot. We conducted four repeated trials of the table-carrying task. Results did not confirm our hypothesis: even though trust increased for non-adaptable participants, a large majority of them remained non-adaptable in the second task as well.

6.1.4.1 Experiment Setting

The task has two parts, each consisting of two trials of task execution. At the end of the first part, we reset the robot belief on participants'

adaptability to a uniform distribution over α. Therefore, in the beginning of the second part, the robot attempted again to guide participants towards the optimal goal, identically to the first part of the task. We recruited participants through Amazon's Mechanical Turk service, using the same inclusion criteria as in section 6.1.2.4. Each participant was compensated $1. Following the data collection process described in section 6.1.2.4, we disregarded participants that had as initial preference the robot goal, resulting in a total of 43 samples. All participants interacted with the robot following the MOMDP policy computed using the proposed formalism. After instructing participants in the task, as well as after each trial, we asked them to rate on a five-point Likert scale their agreement to the following statements.

- "HERB is trustworthy"
- "I am confident in my ability to complete the task"

We used the ratings as direct measurements of participants' self-confidence and trust in the robot.

6.1.4.2 Hypotheses

H4 *The perceived initial robot trustworthiness and the participants' starting self-confidence on their ability to complete the task will have a significant effect on their likelihood to adapt to the robot in the first part of the experiment.* We hypothesized that the more participants trust the robot in the beginning of the task, and the less confident they are on their ability, the more likely they would be to adapt to the robot. In previous work, Lee and Moray [1991] found that control allocation in a supervisory control system is dependent on the difference between the operator's trust of the system and their own self-confidence to control the system under manual control.

H5 *The robot trustworthiness, as perceived by non-adaptable participants, will increase during the first part of the experiment*. We hypothesized that working with a robot that reasons over its estimate on participants' adaptability and changes its own strategy accordingly would increase the non-adaptable participants' trust in the robot. We base this hypothesis by observing in fig. 6.6 that non-adaptable participants in the Mutual-adaptation condition agreed strongly to the statement "HERB is trustworthy" at the end of the task. We focus on non-adaptable participants, since they observe the robot changing its policy to their preference, and previous work has shown that robot adaptation can significantly improve perceived robot trustworthiness [Shah et al., 2011].

H6 *Participants are more likely to follow the robot optimal policy in the second part of the experiment, compared to the first part.* We hypothesized that if, according to hypotheses H4 and H5, trust is associated with increased likelihood of adapting to the robot in the first part of the experiment, and non-adaptable participants trust the robot more after the first part, a significant ratio of these participants would be willing to change their strategy in the second part. Additionally, we expected participants that switched to the robot optimal policy in the first part to continue following that policy in the second part, resulting in an overall increase in the number of subjects that follow the optimal goal.

6.1.4.3 Results and Discussion

We consider Hypothesis **H4**, that the perceived robot trustworthiness and the participants' self-confidence on their ability to complete the task, as measured in the beginning of the experiment, will have a significant effect on their likelihood to adapt to the robot in the first part of the experiment. We performed a logistic regression to ascertain the effects of participants' ratings on these two factors on the likelihood that they adapt to the robot. The logistic regression model was statistically significant $\chi^2(2) = 13.58, p = 0.001$. The model explained 36.2% (Nagelkerke R^2) of the variance in participant's adaptability and correctly classified 74.4% of the cases. Participants that trusted the robot more in the beginning of the task ($\beta = 1.528, p = 0.010$) and were less-confident ($\beta = -1.610, p = 0.008$) were more likely to adapt to the robot in part 1 of the experiment (fig. 6.10, 6.11). This supports hypothesis **H4** of section 6.1.4.2.

Recall Hypothesis **H5**, that the robot trustworthiness, as perceived by non-adaptable participants, will increase during the first part of the experiment. We included in the non-adaptable group all participants that did not change their strategy when working with the robot in the first part of the experiment. The mean estimated adaptability for these participants at the end of the first part was $\hat{\alpha} = 0.16$ [SD = 0.14]. A Wilcoxon signed-rank test indeed showed that non-adaptable participants agreed more strongly that HERB is trustworthy after the first part of the experiment, compared to the beginning of the task ($Z = -3.666, p < 0.001$), as shown in fig. 6.10. In the same figure we see that adaptable participants rated highly their agreement on the robot trustworthiness in the beginning of the task, and their ratings remained relatively similar through the first part of the task. The results above confirm our hypothesis that working with the robot following the MOMDP policy had a significant effect on the non-adaptable participants' trust in the robot.

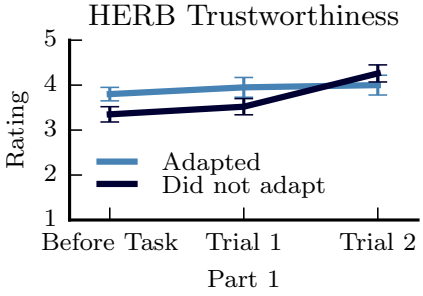


Figure 6.10: Rating of agreement to the statement "HERB is trustworthy." for the first part of the experiment described in section 6.1.4. The two groups indicate participants that adapted / did not adapt to the robot during the first part.

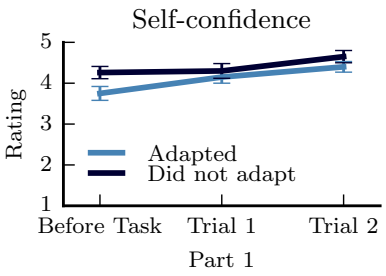


Figure 6.11: Rating of agreement to the statement "I am confident in my ability to complete the task."

To test Hypothesis **H6**, we consider the ratio of participants that followed the robot optimal policy in the first part of the experiment, compared to the second part of the experiment. In the second part, 53% of the participants followed the robot goal, compared to 47% in the first part. A Pearson's chi-square test did not find the difference between the two ratios to be statistically significant ($\chi^2(1, N = 43) = 0.42, p = 0.518$). We observed that all participants that adapted to the robot in the first part, continued following the optimal goal in the second part, as expected. However, only 13% of non-adaptable participants switched strategy in the second part. We observe that even though trust increased for non-adaptable participants, a large majority of them remained non-adaptable in the second task as well. We attribute this to the fact that users, who successfully completed the task in the first part with the robot adapting to their preference, were confident that the same action sequence would result in successful completion in the second part, as well. In fact, a Wilcoxon signed-rank test showed that non-adaptable participants rated their self-confidence on their ability to complete the task significantly higher after the first part, compared to the beginning of the task ($Z = -2.132, p = 0.033$, fig. 6.11). It would be interesting to assess the adaptability of participants after inducing drops in their self-confidence, for instance by providing limited explanation about the task or introducing task "failures," and we leave this for future work.

This experiment showed that non-adaptable participants remained unwilling to adapt to the robot in repeated trials of the same task. Can this result generalize across multiple tasks? This is an important question, since in real-world applications such as home environments, domestic robots are expected to perform a variety of household chores. We conducted a follow-up experiment, where we explored whether the adaptability of participants in one task is informative of their willingness to adapt to the robot at a different task.

6.1.5 Transfer of Adaptability Across Tasks

The previous experiment showed that non-adaptable participants remained unwilling to adapt to the robot in repeated trials of the same task. To test whether this result can generalize across multiple tasks, we conducted an experiment with two different collaborative tasks: a table-carrying task followed by a hallway-crossing task. Results showed that non-adaptable participants in the table-carrying task would be less likely to adapt in the hallway-crossing task.

6.1.5.1 Hallway-Crossing Task

We introduced a new hallway-crossing task, where human and robot cross a hallway (fig. 6.12). As in the table-carrying task, we instructed participants of the task and asked them for their preferred side of the hallway. We then set the same side as the optimal goal for the robot, in order to ensure that the robot optimal policy would conflict with the human preference. The user chose moving actions by clicking on buttons on a user interface (left / right). If human and robot ended up in the same side, a message appeared on the screen notifying the user that they moved in the same direction as the robot. The participant could then choose to remain on that side, or switch sides. The task ended when human and robot ended up in opposite sides of the corridor.

Figure 6.12: Hallway-crossing task. The user faces the robot and can choose to stay in the same side or switch sides. Once the user ends up in the side opposite to the robot's, the task is completed.

6.1.5.2 MOMDP Model of Hallway-Crossing Task

The observable state variables x of the MOMDP formulation were the discretized position of the human and the robot, as well as the human and robot modes for each of the three previous time-steps. We specified two modal policies, each deterministically selecting moving actions towards each side of the corridor. The size of the observable state-space X was 340 states. As in the table-carrying task, we set a history length $k = 3$ and assumed a discrete set of values of the adaptability $\alpha : \{0.0, 0.25, 0.5, 0.75, 1.0\}$. Therefore, the total size of the MOMDP state-space was $5 \times 340 = 1700$ states. The human and robot actions a^H, a^R were deterministic discrete motions towards each side of the corridor. We set the reward function R to be positive at the two goal states based on their relative cost, and 0 elsewhere. We computed the robot policy using the SARSOP solver [Kurniawati et al., 2008].

6.1.5.3 Experiment Setting

We first validated the efficacy of the proposed formalism by doing a user study ($n = 65$) that included only the hallway-crossing task. We recruited participants through Amazon's Mechanical Turk service, using the same inclusion criteria as in section 6.1.2.4. Each participant was compensated \$0.50. 48% of participants adapted to the robot by switching sides, a ratio comparable to that of the table-carrying task experiment (section 6.1.3.2). The mean estimated adaptability for participants that adapted to the robot, which we call "adaptable," was $\hat{\alpha} = 0.85$ [SD = 0.25], and for participants that did not adapt ("non-adaptable") was $\hat{\alpha} = 0.07$ [SD = 0.13].

We then conducted a new human subject experiment, having users

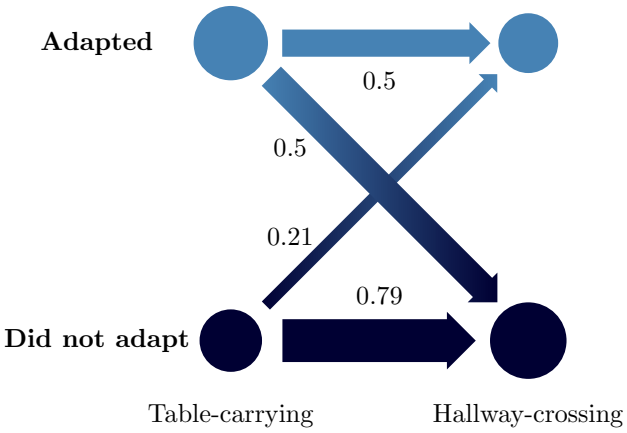


Figure 6.13: Adaptation rate of participants for two consecutive tasks. The lines illustrate transitions, with the numbers indicating transition rates. The thickness of the lines is proportional to the transition rate, whereas the area of the circles is proportional to the number of participants. Whereas 79% of the users that insisted in their strategy in the first task remained non-adaptable in the second task, only 50% of the users that adapted to the robot in the table-carrying task, adapted to the robot in the hallway-crossing task.

do two trials of the table-carrying task described in section 6.1.2.3 (part 1), followed by the hallway-crossing task (part 2). Similarly to the repeated table-carrying task experiment (section 6.1.4), we reset the robot belief on the human adaptability at the end of the first part. We recruited participants through Amazon's Mechanical Turk service, using the same inclusion criteria as in section 6.1.2.4, and following the same data collection process, resulting in a total of $n = 58$ samples. Each participant was compensated $1.30. We make the following hypothesis:

H7 *Participants that did not adapt to the robot in the table-carrying task are less likely to adapt to the robot in the hallway task, compared to participants that changed their strategy in the first task.*

6.1.5.4 Results and Discussion

In line with our hypothesis, a logistic regression model was statistically significant ($\chi^2(1) = 5.30, p = 0.021$), with participants' adaptability in the first task being a significant predictor of their adaptability in the second task ($\beta = 1.335, p = 0.028$). The model explained 11.9% (Nagelkerke R^2) of the variance and correctly classified 62.5% of the cases. The small value of R^2 indicates a weak effect size. Interestingly, whereas 79% of the users that did not adapt to the robot in the first task remained non-adaptable in the second task, only 50% of the users that adapted to the robot in the table-carrying task, adapted to the robot in the hallway task (fig. 6.13).

We interpret this result by observing that all participants that were non-adaptable in the first task saw the robot changing its behavior to their preferred strategy. A large majority expected the robot to behave in the same way in the second task, as well: disagree in the beginning but eventually adapt to their preference, and this encouraged them to insist on their preference also in the second task. In fact, in

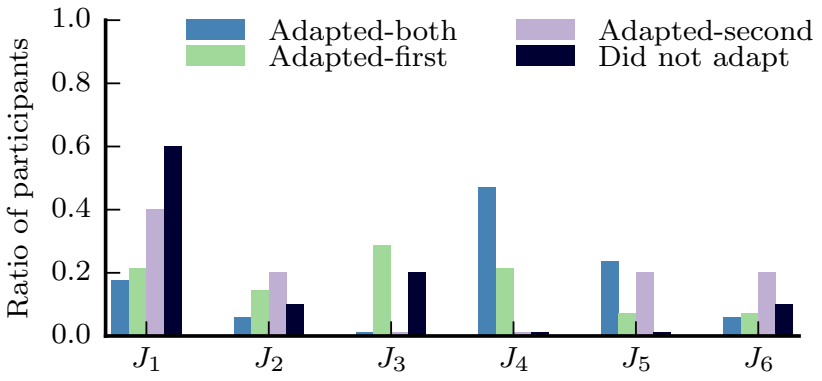


Figure 6.14: Ratio of participants per justification to the total number of participants in each condition. We group the participants based on whether they adapted in both tasks (Adapted-both), in the first [table-carrying] task only (Adapted-first), in the second [hallway-crossing] task only (Adapted-second) and in none of the tasks (Did not adapt).

Table 6.3: Participants' response to question "Did you complete the hallway task following your initial preference? Justify your answer."

	Justification	Example Quote
J1	Expectation on robot behavior	"I knew that the robot would change if I stood my ground"
J2	Simplicity	"I thought it would be easier that I switched"
J3	Task-specific factors	"I was on the correct side (you should walk on the right hand side)"
J4	Robot behavior	"HERB decided to go the same way as I did"
J5	Task completion	"To finish the task in the other end of the hall"
J6	Other	"I tend to stick with my initial choices"

their answers to the open-ended question "Did you complete the hallway task following your initial preference?," they mentioned that "The robot switched in the last [table-carrying] task, and I thought it would this time too", and that "I knew from the table-turning task that HERB would eventually figure it out and move in the opposite direction, so I stood my ground" (J1 in Table 6.3, fig. 6.14). On the other hand, adaptable participants did not have enough information on the robot ability to adapt, since they aligned their own strategy with the robot policy, and they were evenly divided between adaptable and non-adaptable in the second task. 47% of participants that remained adaptable in both tasks attributed the change in their strategy to the robot behavior (J4). Interestingly, 29% of participants that adapted to the robot in the table-carrrying task but insisted on their strategy in the hallway task stated that they did so, "because I was on the correct side (you should walk on the right hand side) and I knew eventually he would move" (J3). We see that task-specific factors, such as social norms, affected the expectation of some participants on the robot adaptability for the hallway task. We hypothesize that

there is an inverse relationship between participants' adaptability, as it evolves over time, and their belief on the robot's own adaptability, and we leave the testing of this hypothesis for future work.

6.1.6 Conclusion

We presented a formalism for human-robot mutual adaptation, which enables guiding the human teammate towards more efficient strategies, while maintaining human trust in the robot. We integrated BAM, a model of human adaptation based on a bounded memory assumption (section 5.1, chapter 5), into a MOMDP formulation, wherein the human adaptability was a partially observable variable. In a human subject experiment ($n = 69$), participants were significantly more likely to adapt to the robot strategy towards the optimal goal when working with a robot utilizing our formalism ($p = 0.036$), compared to cross-training with the robot. Additionally, participants found the performance as a teammate of the robot executing the learned MOMDP policy to be not worse than the performance of the robot that cross-trained with the participants. Finally, the robot was found to be more trustworthy with the learned policy, compared with executing an optimal strategy while ignoring human adaptability ($p = 0.048$). These results indicate that the proposed formalism can significantly improve the effectiveness of human-robot teams, while achieving subjective ratings on robot performance and trust comparable to those of state-of-the-art human-robot team training strategies.

We have shown that BAM can adequately capture human behavior in two collaborative tasks with well-defined task-steps on a relatively fast-paced domain. However, in domains where people typically reflect on a long history of interactions, or on the beliefs of the other agents, such as in a Poker game [Von Neumann and Morgenstern, 2007], people are likely to demonstrate much more complex adaptive behavior. Developing sophisticated predictive models for such domains and integrating them into robot decision making in a principled way, while maintaining computational tractability, is an exciting area for future work.

6.2 Shared-Autonomy

Work done in collaboration with David Hsu and Yu Xiang Zhu.

Assistive robot arms show great promise in increasing the independence of people with upper extremity disabilities [Hillman et al., 2002, Prior, 1990, Sijs et al., 2007]. However, when a user teleoperates directly a robotic arm via an interface such as a joystick, the limitation of the interface, combined with the increased capability and complexity of robot arms, often makes it difficult or tedious to

Figure 6.15: The user guides the robot towards an unstable grasp, resulting in task failure.

accomplish complex tasks.

Shared autonomy alleviates this issue by combining direct teleoperation with autonomous assistance [Kofman et al., 2005, Dragan and Srinivasa, 2013b, Yu et al., 2005, Trautman, 2015, Gopinath et al., 2017]. In recent work by Javdani et al. [2015], the robot estimates a distribution of user goals based on the history of user inputs, and assists the user for that distribution. The user is assumed to be always right about their goal choice. Therefore, if the assistance strategy knows the user's goal, it will select actions to minimize the cost-to-go to that goal. This assumption is often not true, however. For instance, a user may choose an unstable grasp when picking up an object (fig. 6.15), or they may arrange items in the wrong order by stacking a heavy item on top of a fragile one. Fig. 6.16 shows a shared autonomy scenario, where the user teleoperates the robot towards the left bottle. We assume that the robot *knows the optimal goal for the task*: picking up the right bottle is a better choice, for instance because the left bottle is too heavy, or because the robot has less uncertainty about the right bottle's location. Intuitively, if the human insists on the left bottle, the robot should comply; failing to do so can have a negative effect on the user's trust in the robot, which may lead to disuse of the system [Hancock et al., 2011, Salem et al., 2015, Lee et al., 2013]. If the human is willing to adapt by aligning its actions with the robot, which has been observed in adaptation between humans and artifacts [Xu et al., 2009, Komatsu et al., 2005], the robot should insist towards the optimal goal. The human-robot team then exhibits *a mutually adaptive behavior, where the robot adapts its own actions by reasoning over the adaptability of the human teammate*.

In section 6.1, we proposed a human-robot mutual adaptation formalism for a shared location collaboration task. In this section, *we generalize the formalism* for the shared-autonomy setting.

We identify that in the shared-autonomy setting (1) tasks may typically exhibit less structure than in the collaboration domain, which limits the observability of the user's intent, and (2) only robot actions directly affect task progress. We address the first challenge by including the human mode m^{H} as an additional latent variable in a mixed-observability Markov decision process (MOMDP) [Ong et al.,

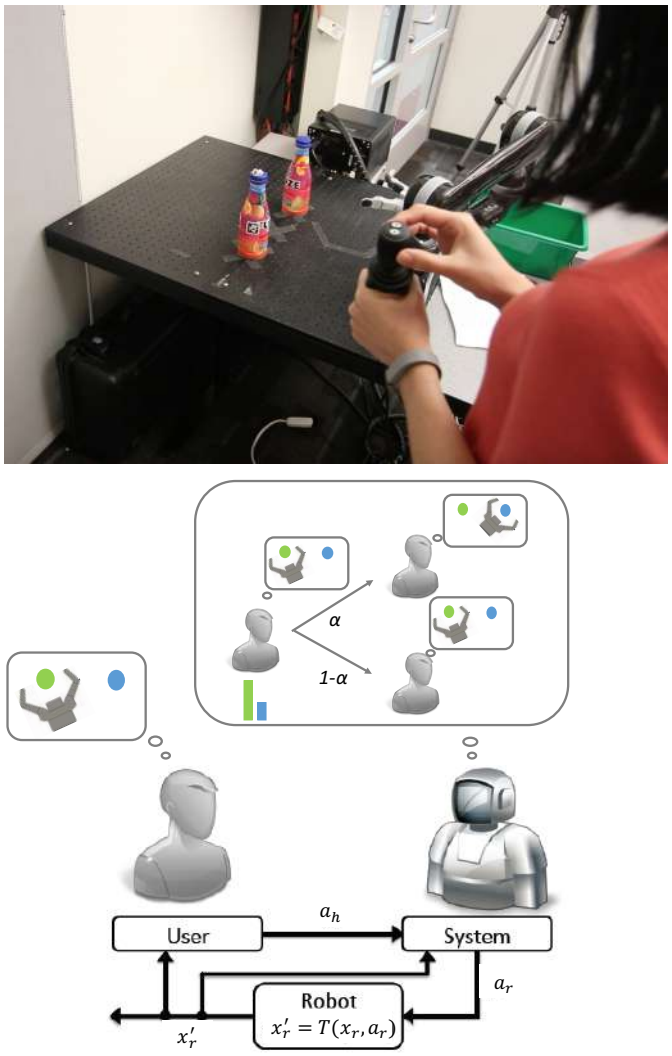


Figure 6.16: Table clearing task in a shared autonomy setting. The user operates the robot using a joystick interface and moves the robot towards the left bottle, which is a suboptimal goal. The robot plans its actions based on its estimate of the current human goal and the probability α of the human switching towards a new goal indicated by the robot.

2010]. This allows the robot to maintain a probability distribution over the user goals based on the history of operator inputs. We also take into account the uncertainty that the human has on the robot goal by modeling the human as having a probability distribution over the robot modes m^{R} (section 6.2.1). We address the second challenge by treating the human actions as observations that do not affect the world state. This allows the robot to infer simultaneously the human mode m^{H} and the human adaptability α, reason over how likely the human is to switch their goal based on the robot actions, and guide the human towards the optimal goal while retaining their trust.

We conducted a human subject experiment ($n = 51$) with an assistive robotic arm on a table-clearing task. Results show that the proposed formalism significantly improved human-robot team per-

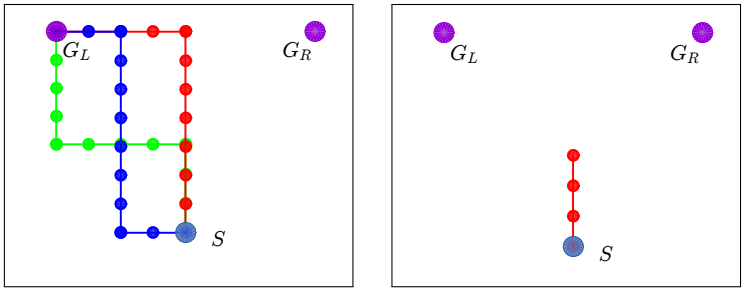


Figure 6.17: (left) Paths corresponding to three different modal policies that lead to the same goal G_L. We use a stochastic modal policy m_L to compactly represent all feasible paths from S to G_L. (right) The robot moving upwards from point S could be moving towards either G_L or G_R.

formance, compared to the robot following participants' preference, while retaining a high level of human trust in the robot.

6.2.1 *Human and Robot Mode Inference*

When the human provides an input through a joystick interface, the robot makes an inference on the human mode. In the example table-clearing task of fig. 6.16, if the robot moves to the right, the human will infer that the robot follows a modal policy towards the right bottle. Similarly, if the human moves the joystick to the left, the robot will consider more likely that the human follows a modal policy towards the left bottle. In this section, we formalize the inference that human and robot make on each other's goals.

6.2.1.1 Stochastic Modal Policies

In the shared autonomy setting, there can be a very large number of modal policies that lead to the same goal. We use as example the table-clearing task of fig. 6.16. We let G_L represent the left bottle, G_R the right bottle, and S the starting end-effector position of the robot. fig. 6.17-left shows paths from three different modal policies that lead to the same goal G_L. Accounting for a large set of modes can increase the computational cost, in particular if we assume that the human mode is partially observable (section 6.2.3.1).

In section 5.1, chapter 5, we defined a modal policy as a stochastic joint-policy over human and robot actions, so that $m\colon X^{\mathrm{R}} \times H_t \rightarrow \Pi(A^{\mathrm{R}}) \times \Pi(A^{\mathrm{H}})$. A stochastic modal policy compactly represents a probability distribution over paths and allows us to reason probabilistically about the future actions of an agent that does not move in a perfectly predictable manner. For instance, we can use the principle of maximum entropy to create a probability distribution over all paths from start to the goal [Ziebart et al., 2009, 2008]. While a stochastic modal policy represents the uncertainty of the observer over paths, we do not require the agent to actually follow a stochastic policy.

6.2.1.2 Full Observability Assumption

While m^R, m^H can be assumed to be observable for a variety of structured tasks in the collaboration domain(section 6.1), this is not the case for the shared autonomy setting because of the following factors:
Different policies invoke the same action. Assume two modal policies in fig. 6.17, one for the left goal shown in red in fig. 6.17-left, and a symmetric policy for the right goal (not shown). An agent moving upwards (fig. 6.17-right) could be following either of the two with equal probability. In that case, inference of the exact modal policy without any prior information is impossible, and the observer needs to maintain a uniform belief over the two policies.
Human inputs are noisy. The user provides its inputs to the system through a joystick interface. These inputs are noisy: the user may "overshoot" an intended path and correct their input, or move the joystick in the wrong direction. In the fully observable case, this would result in an incorrect inference of the human mode. Maintaining a belief over modal policies allows robustness to human mistakes.

This leads us to assume that modal policies are *partially observable.* We model how the human infers the robot mode, as well as how the robot infers the human mode, in the following sections.

6.2.1.3 Robot Mode Inference

The bounded memory assumption (section 5.1.1, chapter 5) dictates that the human does not recall the whole history of states and actions, but only a recent history of the last k time-steps. The human attributes the robot actions to a robot mode m^R.

$$\begin{aligned} P(m^R|h_k, x_t^R, a_t^R) &= P(m^R|x_{t-k+1}^R, a_{t-k+1}^R, ..., x_t^R, a_t^R) \\ &= \eta\ P(a_{t-k+1}^R, ..., a_t^R|m^R, x_{t-k+1}^R, ..., x_t^R) \end{aligned} \tag{6.3}$$

We consider modal policies that generate actions based only on the current world state, so that $M : X^R \to \Pi(A^H) \times \Pi(A^R)$.

Therefore eq. 6.3 can be simplified as follows, where $m^R(x_t^R, a_t^R)$ denotes the probability of the robot taking action a^R at time t, if it follows modal policy m^R:

$$P(m^R|h_k, x_t^R, a_t^R) = \eta\ m^R(x_{t-k+1}^R, a_{t-k+1}^R)...m^R(x_t^R, a_t^R) \tag{6.4}$$

$P(m^R|h_k, x_t^R, a_t^R)$ is the [estimated by the robot] human belief on the robot mode m^R.

6.2.1.4 Human Mode Inference

To infer the human mode, we need to implement the dynamics model T_{m^H} that describes how the human mode evolves over time.

Additionally, contrary to the collaboration setting, the human inputs do not affect directly the world state. Instead, the robot uses them as observations, based on an observation function O, in order to update its belief on the human mode.

In section 3 we defined a transition function $T_{m^{\mathrm{H}}}$, that indicates the probability of the human switching from mode m^{H} to a new mode m'^{H}, given a history h_k and their adaptability α. We simplify the notation, so that $x^{\mathrm{R}} \equiv x_t^{\mathrm{R}}$, $a^{\mathrm{R}} \equiv a_t^{\mathrm{R}}$ and $x \equiv (h_k, x^{\mathrm{R}})$:

$$\begin{aligned} &T_{m^{\mathrm{H}}}(x, \alpha, m^{\mathrm{H}}, a^{\mathrm{R}}, m'^{\mathrm{H}}) = P(m'^{\mathrm{H}} | x, \alpha, m^{\mathrm{H}}, a^{\mathrm{R}}) \\ &= \sum_{m^{\mathrm{R}}} P(m'^{\mathrm{H}}, m^{\mathrm{R}} | x, \alpha, m^{\mathrm{H}}, a^{\mathrm{R}}) \\ &= \sum_{m^{\mathrm{R}}} P(m'^{\mathrm{H}} | x, \alpha, m^{\mathrm{H}}, a^{\mathrm{R}}, m^{\mathrm{R}}) \times P(m^{\mathrm{R}} | x, \alpha, m^{\mathrm{H}}, a^{\mathrm{R}}) \\ &= \sum_{m^{\mathrm{R}}} P(m'^{\mathrm{H}} | \alpha, m^{\mathrm{H}}, m^{\mathrm{R}}) \times P(m^{\mathrm{R}} | x, a^{\mathrm{R}}) \end{aligned} \tag{6.5}$$

The first term gives the probability of the human switching to a new mode m'^{H}, if the human mode is m^{H} and the robot mode is m^{R}. Based on the BAM model (section 5.1, chapter 5), the human switches to m^{R}, with probability α and stays at m^{H} with probability $1 - \alpha$, where α is the human adaptability. If $\alpha = 1$, the human switches to m^{R} with certainty. If $\alpha = 0$, the human insists on their mode m^{H} and does not adapt. Therefore:

$$P(m'^{\mathrm{H}} | \alpha, m^{\mathrm{H}}, m^{\mathrm{R}}) = \begin{cases} \alpha & m'^{\mathrm{H}} \equiv m^{\mathrm{R}} \\ 1 - \alpha & m'^{\mathrm{H}} \equiv m^{\mathrm{H}} \\ 0 & \text{otherwise} \end{cases} \tag{6.6}$$

The second term in eq. 6.5 is computed using eq. 6.4, and it is the [estimated by the human] robot mode.

Eq. 6.5 describes that the probability of the human switching to a new robot mode m^{R} depends on the human adaptability α, as well as on the uncertainty that the human has about the robot following m^{R}. This allows the robot to compute the probability of the human switching to the robot mode, given each robot action.

The observation function $O \colon X^{\mathrm{R}} \times M \to \Pi(A^{\mathrm{H}})$ defines a probability distribution over human actions a^{H}. This distribution is specified by the stochastic modal policy $m^{\mathrm{H}} \in M$. Given the above, the human mode m^{H} can be estimated by a Bayes filter, with $b(m^{\mathrm{H}})$ the robot's previous belief on m^{H}:

$$b'(m'^{\mathrm{H}}) = \eta \; O(m'^{\mathrm{H}}, x'^{\mathrm{R}}, a^{\mathrm{H}}) \sum_{m^{\mathrm{H}} \in M} T_{m^{\mathrm{H}}}(x, \alpha, m^{\mathrm{H}}, a^{\mathrm{R}}, m'^{\mathrm{H}}) b(m^{\mathrm{H}}) \tag{6.7}$$

In this section, we assumed that α is known to the robot. In practice, the robot needs to estimate both m^{H} and α. We formulate this in section 6.2.3.1.

6.2.2 *Disagreement between Modes*

In the previous section we formalized the inference that human and robot make on each other's goals. Based on that, the robot can infer the human goal and it can reason over how likely the human is to switch goals given a robot action.

Intuitively, if the human insists on their goal, the robot should follow the human goal, even if it is suboptimal, in order to retain human trust. If the human is willing to change goals, the robot should move towards the optimal goal. We enable this behavior by proposing in the robot's reward function a penalty for disagreement between human and robot modes. The intuition is that if the human is non-adaptable, they will insist on their own mode throughout the task, therefore the expected accumulated cost of disagreeing with the human will outweigh the reward of the optimal goal. In that case, the robot will follow the human preference. If the human is adaptable, the robot will move towards the optimal goal, since it will expect the human to change modes.

As described in the section 5.1.5 of chapter 5, we formulate the reward function that the robot is maximizing, so that there is a penalty for following a mode that is perceived to be different than the human's mode. We assume a set of goal states G:

$$R(x, m^{\mathrm{H}}, a^{\mathrm{R}}) = \begin{cases} R_{goal} & : x^{\mathrm{R}} \in G \\ R_{other} & : x^{\mathrm{R}} \notin G \end{cases} \tag{6.8}$$

If the robot is at a goal state $x^{\mathrm{R}} \in G$, a positive reward associated with that goal is returned, regardless of the human mode m^{H} and robot mode m^{R}. Otherwise, there is a penalty $C < 0$ for disagreement between m^{H} and m^{R}, induced in R_{other}. The human does not observe m^{R} directly, but estimates it from the recent history of robot states and actions (section 6.2.1.3). Therefore, R_{other} is computed so that the penalty for disagreement is weighted by the [estimated by the human] probability of the robot actually following m^{R}:

$$R_{other} = \sum_{m^{\mathrm{R}}} R_m(m^{\mathrm{H}}, m^{\mathrm{R}}) P(m^{\mathrm{R}} | x, a^{\mathrm{R}}) \tag{6.9}$$

$$\text{where } R_m(m^{\mathrm{H}}, m^{\mathrm{R}}) = \begin{cases} 0 & : m^{\mathrm{H}} \equiv m^{\mathrm{R}} \\ C & : m^{\mathrm{H}} \neq m^{\mathrm{R}} \end{cases} \tag{6.10}$$

6.2.3 Robot Planning

6.2.3.1 MOMDP Formulation

In section 6.2.1.4, we showed how the robot estimates the human mode, and how it computes the probability of the human switching to the robot mode based on the human adaptability. In section 6.2.2, we defined a reward function that the robot is maximizing, which captures the trade-off between going to the optimal goal and following the human mode. Both the human adaptability and the human mode are not directly observable. Therefore, the robot needs to estimate them through interaction, while performing the task. This leads us to formulate this problem as a mixed-observability Markov Decision Process (MOMDP) [Ong et al., 2010]. This formulation allows us to compute an optimal policy for the robot that will maximize the expected reward that the human-robot team will receive, given the robot's estimates of the human adaptability and of the human mode. We define a MOMDP as a tuple $\{X, Y, A^{\mathrm{R}}, \mathcal{T}_x, \mathcal{T}_\alpha, \mathcal{T}_{m^{\mathrm{H}}}, R, \Omega, O\}$:

- $X : X^{\mathrm{R}} \times H_K$ is the set of observable variables. These are the current robot configuration x^{R}, as well as the history h_k. Since x^{R} transitions deterministically, we only need to register the current robot state and robot actions $a^{\mathrm{R}}_{t-k+1}, ..., a^{\mathrm{R}}_t$. We assume that the set of world states X^{w} is identical to the set of robot configurations X^{R}.
- $Y : \mathcal{A} \times M$ is the set of partially observable variables. These are the human adaptability $\alpha \in A$, and the human mode $m^{\mathrm{H}} \in M$.
- A^{R} is a finite set of robot actions. We model actions as transitions between discrete robot configurations.
- $\mathcal{T}_x : X \times A^{\mathrm{R}} \longrightarrow X$ is a deterministic mapping from a robot configuration x^{R}, history h_k and action a^{R}, to a subsequent configuration x'^{R} and history h'_k.
- $\mathcal{T}_\alpha : \mathcal{A} \times A^{\mathrm{R}} \longrightarrow \Pi(\mathcal{A})$ is the probability of the human adaptability being α' at the next time step, if the adaptability of the human at time t is α and the robot takes action a^{R}. We assume the human adaptability to be fixed throughout the task.
- $\mathcal{T}_{m^{\mathrm{H}}} : X \times \mathcal{A} \times M \times A^{\mathrm{R}} \longrightarrow \Pi(M)$ is the probability of the human switching from mode m^{H} to a new mode m'^{H}, given a history h_k, robot state x^{R}, human adaptability α and robot action a^{R}. It is computed using eq. 6.5, section 6.2.1.4.
- $R : X \times M \times A^{\mathrm{R}} \longrightarrow \mathbb{R}$ is a reward function that gives an immediate reward for the robot taking action a^{R} given a history h_k, human mode m^{H} and robot state x^{R}. It is defined in eq. 6.8, section 6.2.2.

- Ω is the set of observations that the robot receives. An observation is a human input $a^{\mathrm{H}} \in A^{\mathrm{H}}$ ($\Omega \equiv A^{\mathrm{H}}$).
- $O : M \times X^{\mathrm{R}} \longrightarrow \Pi(\Omega)$ is the observation function, which gives a probability distribution over human actions for a mode m^{H} at state x^{R}. This distribution is specified by the stochastic modal policy $m^{\mathrm{H}} \in M$.

6.2.4 *Belief Update*

Based on the above, the belief update for the MOMDP is:

$$b'(\alpha', m'^{\mathrm{H}}) = \eta O(m'^{\mathrm{H}}, x'^{\mathrm{R}}, a^{\mathrm{H}}) \sum_{\alpha \in \mathcal{A}} \sum_{m^{\mathrm{H}} \in M} \mathcal{T}_x(x, a^{\mathrm{R}}, x') \mathcal{T}_\alpha(\alpha, a^{\mathrm{R}}, \alpha') \mathcal{T}_{m^{\mathrm{H}}}(x, \alpha, m^{\mathrm{H}}, a^{\mathrm{R}}, m'^{\mathrm{H}}) b(\alpha, m^{\mathrm{H}}) \tag{6.11}$$

We note that since the MOMDP has two partially observable variables, α and m^{H}, the robot maintains a joint probability distribution over both variables.

6.2.5 *Robot Policy*

We solve the MOMDP for a robot policy that is optimal with respect to the robot's expected total reward.

The stochastic modal policies may assign multiple actions at a given state. Therefore, even if $m^{\mathrm{H}} \equiv m^{\mathrm{R}}$, a^{R} may not match the human input a^{H}. Such disagreements are unnecessary when human and robot modes are the same. Therefore, we let the robot actions match the human inputs, if the robot has enough confidence that robot and human modes (computed using eq. 6.4, 6.7) are identical in the current time-step. Otherwise, the robot executes the action specified by the MOMDP optimal policy. We leave for future work adding a penalty for disagreement between actions, which we hypothesize it would result in similar behavior.

6.2.6 *Simulations*

Fig. 6.18 shows the robot behavior for two simulated users, one with low adaptability (User 1, $\alpha = 0.0$), and one with high adaptability (User 2, $\alpha = 0.75$) for a shared autonomy scenario with two goals, G_L and G_R, corresponding to modal policies m_L and m_R respectively. Both users start with modal policy m_L (left goal). The human and robot actions are {move-left, move-right, move-forward}. The robot uses the human input to estimate both m^{H} and α. For both users, the upper row plots the robot trajectory (red dots), the human input (green arrow) and the robot action (gray arrow)

over time. The middle row plots the estimate of α over time, where $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. Each graph plots α versus the probability of α. The lower row plots $m \in \{m_L, m_R\}$ versus the probability of m. Columns indicate successive time-steps. User 1 insists on their initial strategy throughout the task and the robot complies, whereas User 2 adapts to the robot and ends up following m_R. We set a bounded memory of $k = 1$ time-step. If human and robot disagree and the human insists on their modal policy, then the MOMDP belief is updated so that smaller values of adaptability α have higher probability (lower adaptability). It the human aligns its inputs to the robot mode, larger values become more likely. If the robot infers the human to be adaptable, it moves towards the optimasl goal. Otherwise, it complies with the human, thus retaining their trust.

Fig. 6.19 shows the team-performance over α, averaged over 1000 runs with simulated users. We evaluate performance by the reward of the goal achieved, where R_{opt} is the reward for the optimal and R_{sub} for the sub-optimal goal. We see that the more adaptable the user, the more often the robot will reach the optimal goal. Additionally, we observe that for $\alpha = 0.0$, the performance is higher than R_{sub}. This is because the simulated user may choose to move forward in the first time-steps; when the robot infers that they are stubborn, it is already close to the optimal goal and continues moving to that goal.

6.2.7 Human Subject Experiment

We conduct a human subject experiment ($n = 51$) in a shared autonomy setting. We are interested in showing that the human-robot mutual adaptation formalism can improve the performance of human-robot teams, while retaining high levels of perceived collaboration and trust in the robot in the shared autonomy domain.

On one extreme, we "fix" the robot policy, so that the robot always moves towards the optimal goal, ignoring human adaptability. We hypothesize that this will have a negative effect on human trust and perceived robot performance as a teammate. On the other extreme, we have the robot assist the human in achieving their desired goal.

We show that the proposed formalism achieves a trade-off between the two: when the human is non-adaptable, the robot follows the human preference. Otherwise, the robot insists on the optimal way of completing the task, leading to significantly better policies, compared to following the human preference, while achieving a high level of trust.

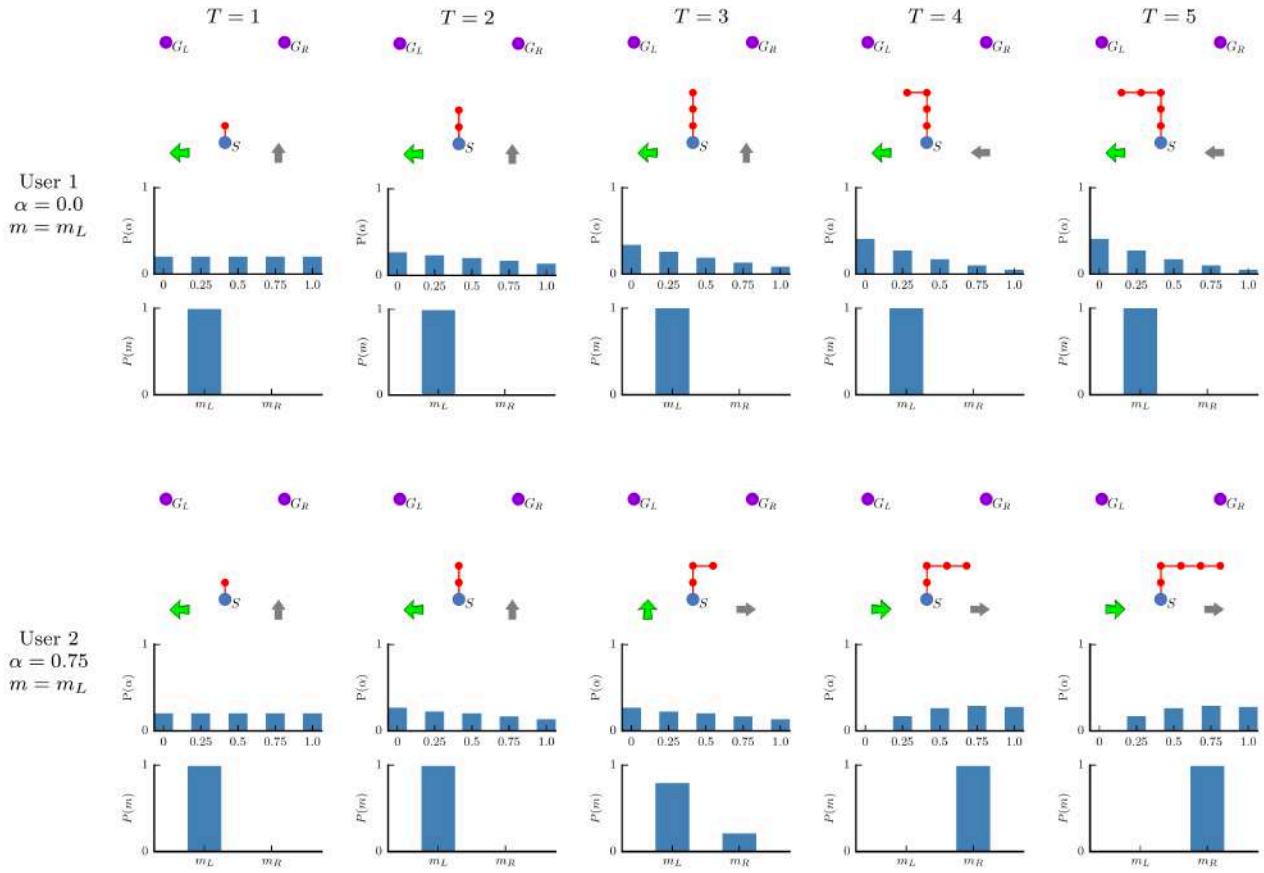


Figure 6.18: Sample runs on a shared autonomy scenario with two goals G_L,G_R and two simulated humans of adaptability level α=0 and 0.75.

6.2.7.1 Independent Variables

No-adaptation session. The robot executes a fixed policy, always acting towards the optimal goal.
Mutual-adaptation session. The robot executes the MOMDP policy of section 6.2.5.
One-way adaptation session. The robot estimates a distribution over user goals, and adapts to the user following their preference, assisting them for that distribution [Javdani et al., 2015]. We compute the robot policy in that condition by fixing the adaptability value to 0 in our model and assigning equal reward to both goals.

6.2.7.2 Hypotheses

H1 *The performance of teams in the No-adaptation condition will be better than of teams in the Mutual-adaptation condition, which will in turn be better than of teams in the One-way adaptation condition.* We expected teams

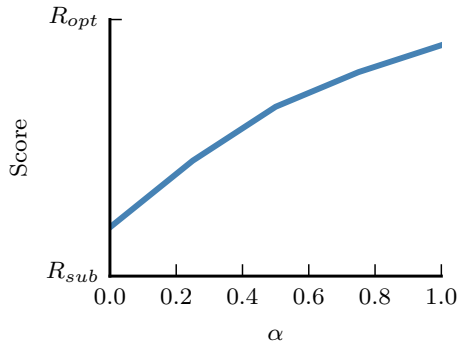


Figure 6.19: Mean performance for simulated users of different adaptability α.

in the No-adaptation condition to outperform the teams in the other conditions, since the robot will always go to the optimal goal. In the Mutual-adaptation condition, we expected a significant number of users to adapt to the robot and switch their strategy towards the optimal goal. Therefore, we posited that this would result in an overall higher reward, compared to the reward resulting from the robot following the participants' preference throughout the task (One-way adaptation).

H2 *Participants that work with the robot in the One-way adaptation condition will rate higher their trust in the robot, as well as their perceived collaboration with the robot, compared to working with the robot in the Mutual-adaptation condition,. Additionally, participants in the Mutual-adaptation condition will give higher ratings, compared to working with the robot in the No-adaptation condition.* We expected users to trust the robot more in the One-way adaptation condition than in the other conditions, since in that condition the robot will always follow their preference. In the Mutual-adaptation condition, we expected users to trust the robot more and perceive it as a better teammate, compared with the robot that executed a fixed strategy ignoring users' adaptability (No-adaptation). Previous work in collaborative tasks has shown a significant improvement in human trust, when the robot had the ability to adapt to the human parter [Shah et al., 2011, Lasota and Shah, 2015].

6.2.7.3 Experiment Setting: A Table Clearing Task

Participants were asked to clear a table off two bottles placed symmetrically, by providing inputs to a robotic arm through a joystick interface (fig. 6.16). They controlled the robot in Cartesian space by moving it in three different directions: left, forward and right. We first instructed them in the task, and asked them to do two training sessions, where they practiced controlling the robot with the joystick. We then asked them to choose which of the two bottles they would

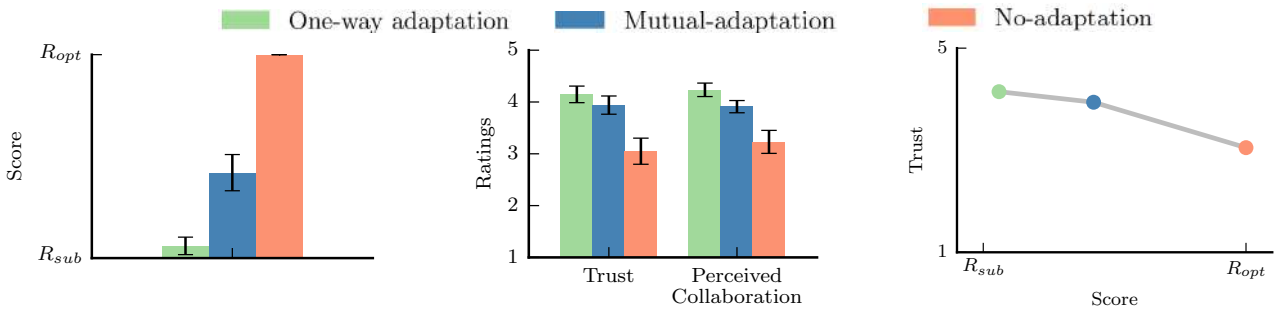


Figure 6.20: Findings for objective and subjective measures.

like the robot to grab first, and we set the robot policy, so that the other bottle was the optimal goal. This emulates a scenario where, for instance, the robot would be unable to grasp one bottle without dropping the other, or where one bottle would be heavier than the other and should be placed in the bin first. In the one-way and mutual adaptation conditions, we told them that "the robot has a mind of its own, and it may choose not to follow your inputs." Participants then did the task three times in all conditions, and then answered a post-experimental questionnaire that used a five-point Likert scale to assess their responses to working with the robot. Additionally, in a video-taped interview at the end of the task, we asked participants that had changed strategy during the task to justify their action.

6.2.7.4 Subject Allocation

We recruited 51 participants from the local community, and chose a between-subjects design in order to not bias the users with policies from previous conditions.

6.2.7.5 MOMDP Model

The size of the observable state-space X was 52 states. We empirically found that a history length of $k = 1$ in BAM was sufficient for this task, since most of the subjects that changed their preference did so reacting to the previous robot action. The human and robot actions were {move-left, move-right, move-forward}. We specified two stochastic modal policies $\{m_L, m_R\}$, one for each goal. We additionally assumed a discrete set of values of the adaptability α : $\{0.0, 0.25, 0.5, 0.75, 1.0\}$. Therefore, the total size of the MOMDP state-space was $5 \times 2 \times 52 = 520$ states. We selected the reward so that $R_{opt} = 11$ for the optimal goal, $R_{sub} = 10$ for the suboptimal goal, and $C = -0.32$ for the cost of mode disagreement (eq. 6.10). We computed the robot policy using the SARSOP solver [Kurniawati et al., 2008], a point-based approximation algorithm which, combined with

the MOMDP formulation, can scale up to hundreds of thousands of states [Bandyopadhyay et al., 2013].

6.2.8 *Analysis*

6.2.8.1 Objective Measures

We consider hypothesis **H1**, that the performance of teams in the No-adaptation condition will be better than of teams in the Mutual-adaptation condition, which in turn will be better than of teams in the One-way adaptation condition.

Nine participants out of 16 (56%) in the Mutual-adaptation condition guided the robot towards the optimal goal, which was different than their initial preference, during the final trial of the task, while 12 out of 16 (75%) did so at one or more of the three trials. From the participants that changed their preference, only one stated that they did so for reasons irrelevant to the robot policy. On the other hand, only two participants out of 17 in the One-way adaptation condition changed goals during the task, while 15 out of 17 guided the robot towards their preferred, suboptimal goal in all trials. This indicates that the adaptation observed in the Mutual-adaptation condition was caused by the robot behavior.

We evaluate team performance by computing the mean reward over the three trials, with the reward for each trial being R_{opt} if the robot reached the optimal goal and R_{sub} if the robot reached the suboptimal goal (fig. 6.20-left). As expected, a Kruskal-Wallis H test showed that there was a statistically significant difference in performance among the different conditions ($\chi^2(2) = 39.84, p < 0.001$). Pairwise two-tailed Mann-Whitney-Wilcoxon tests with Bonferroni corrections showed the difference to be statistically significant between the No-adaptation and Mutual-adaptation ($U = 28.5, p < 0.001$), and Mutual-adaptation and One-way adaptation ($U = 49.5, p = 0.001$) conditions. This supports our hypothesis.

6.2.8.2 Subjective Measures

Recall hypothesis **H2**, that participants in the Mutual-adaptation condition would rate their trust and perceived collaboration with the robot higher than in the No-adaptation condition, but lower than in the One-way adaptation condition. Table I shows the two subjective scales that we used. The *trust* scales were used as-is from Hoffman [2013]. We additionally chose a set of questions related to participants' *perceived collaboration* with the robot.

Both scales had good consistency. Scale items were combined into a score. Fig. 6.20-center shows that both participants' trust ($M =$

3.94, $SE = 0.18$) and perceived collaboration ($M = 3.91, SE = 0.12$) were high in the Mutual-adaptation condition. One-way ANOVAs showed a statistically significant difference between the three conditions in both trust ($F(2,48) = 8.370, p = 0.001$) and perceived collaboration ($F(2,48) = 9.552, p < 0.001$). Tukey post-hoc tests revealed that participants of the Mutual-adaptation condition trusted the robot more, compared to participants that worked with the robot in the No-adaptation condition ($p = 0.010$). Additionally, they rated higher their perceived collaboration with the robot ($p = 0.017$). However, there was no significant difference in either measure between participants in the One-way adaptation and Mutual-adaptation conditions. We attribute these results to the fact that the MOMDP formulation allowed the robot to reason over its estimate of the adaptability of its teammate; if the teammate insisted towards the suboptimal goal, the robot responded to the input commands and followed the user's preference. If the participant changed their inputs based on the robot actions, the robot guided them towards the optimal goal, while retaining a high level of trust. By contrast, the robot in the No-adaptation condition always moved towards the optimal goal ignoring participants' inputs, which in turn had a negative effect on subjective measures.

6.2.9 *Sensitivity Analysis*

We want to explore how sensitive is the robot performance to the value of the mode disagreement penalty C. We vary the value of C and simulate the task execution over α, averaged over 10000 runs with simulated users. Similarly to section 6.2.6, we evaluate performance by the reward of the goal achieved, where R_{opt} is the reward for the optimal and R_{sub} for the sub-optimal goal.

Fig. 6.21 shows the team performance for different values of C. $C = -0.32$ is the value of the cost that we used for the simulations in section 6.2.6 and the experiments in section 6.2.7. We observe that decreasing the magnitude of the cost ($C > 0.32$) results in the optimal performance regardless of α. This is because the robot always ignores the user and goes towards the optimal goal; the robot policy becomes identical to the one in the No-adaptation session. On the other hand, increasing the magnitude of cost ($C < 0.32$) results in lower values in the y-axis, since the robot becomes more reluctant to disagree with the human user. Finally, for $C \leq 0.45$, the performance does not change and becomes close to R_{sub} for any α. A small increase in performance for higher values of α for the $C \leq 0.45$ curve occurs because, even though the robot follows the human mode, the human may misinterpret a robot forward action as an action towards the

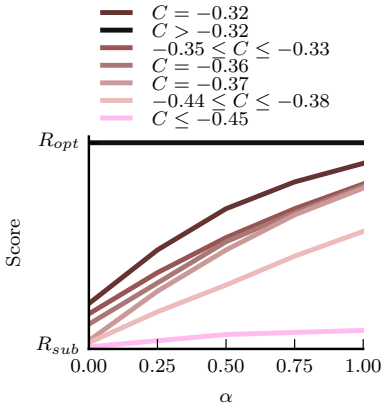


Figure 6.21: Mean performance for simulated users and robot policies of varying mode disagreement cost C

right goal, and adapt to that with probability α.

We observe that the robot policy is particularly sensitive to the values of C. One solution would be to vary C across the state-space of the task, rather than assuming a constant penalty for mode-disagreement. Alternatively, rather than implicitly penalizing disagreement as a way to retain human trust, we could model trust as an additional latent variable in the MOMDP, and include it as a reward function parameter. We leave this for future investigation.

6.2.10 Conclusion

In this work, we proposed a human-robot mutual adaptation formalism in a shared autonomy setting. In a human subject experiment, we compared the policy computed with our formalism, with an assistance policy, where the robot helped participants to achieve their intended goal, and with a fixed policy where the robot always went towards the optimal goal.

As fig. 6.20 illustrates, participants in the one-way adaptation condition had the worst performance, since they guided the robot towards a suboptimal goal. The fixed policy achieved maximum performance, as expected. However, this came to the detriment of human trust in the robot. On the other hand, the assistance policy in the One-way adaptation condition resulted in the highest trust ratings — albeit not significantly higher than the ratings in the Mutual-adaptation condition — since the robot always followed the user preference and there was no goal disagreement between human and robot. Mutual-adaptation balanced the trade-off between optimizing performance and retaining trust: users in that condition trusted the robot more than in the No-adaptation condition, and performed better than in the One-way adaptation condition.

Fig. 6.20-right shows the three conditions with respect to trust and performance scores. We can make the MOMDP policy identical to either of the two policies in the end-points, by changing the MOMDP model parameters. If we fix in the model the human adaptability to 0 and assign equal costs for both goals, the robot would assist the user in their goal (One-way adaptation). If we fix adaptability to 1 in the model (or we remove the penalty for mode disagreement), the robot will always go to the optimal goal (fixed policy).

The presented table-clearing task can be generalized without significant modifications to tasks with a large number of goals, human inputs and robot actions, such as picking good grasps in manipulation tasks (fig. 6.15): The state-space size increases linearly with $(1/dt)$, where dt a discrete time-step, and with the number of modal policies. On the other hand, the number of observable states is poly-

TABLE I: SUBJECTIVE MEASURES

Trust $\alpha = .85$
1.I trusted the robot to do the right thing at the right time.
2.The robot was trustworthy.
Perceived Collaboration $\alpha = .81$
1.I was satisfied with ADA and my performance.
2.The robot and I worked towards mutually agreed upon goals.
3.The robot and I collaborated well together
4.The robot's actions were reasonable.
5.The robot was responsive to me.

nomial to the number of robot actions ($O(|A^R|^k)$), since each state includes history h_k: For tasks with large $|A^R|$ and memory length k, we could approximate h_k using feature-based representations.

6.3 Discussion

In this chapter, we relaxed the assumption of a known human type θ. Instead, we treated θ as a latent variable in a partially observable stochastic process; this allowed the robot to take information seeking actions to infer online the human type θ. The human type informs how the human adapts to the robot. This results in human-robot *mutual adaptation*. The robot adapts its own actions, by building online a model of human adaptation to the robot.

We are excited to have brought about a better understanding of the relationships between adaptability performance and trust in collaboration and shared-autonomy settings. In particular, we have showed that the mutual adaptation formalism significantly improved the performance of human-robot teams.

So far, we have considered that the human adaptability is constant throughout the task. In other words, we have assumed that if a user is non-adaptable, this does not change as they interact with the robot. In the next chapter, we relax this assumption by introducing *verbal communication* from the robot to the human, and we investigate how different types of utterances affect team performance and user trust in the robot.

7
Mutual Adaptation with Verbal Communication

Work done in collaboration with Jodi Forlizzi and Minae Kwon.

This chapter generalizes the mutual-adaptation formalism of chapter 6 to include verbal communication. Our generalized formalism [1] enables a robot to *combine optimally verbal communication and actions towards task completion* to guide a human teammate towards a better way of doing a collaborative task.

To demonstrate the applicability of the formalism, we revisit the table-carrying task of chapter 5 (Fig. 7.1). We focus on the robot verbally communicating two types of information: *how* the robot wants them to behave, and *why* the robot is behaving this way. Therefore, we identify two types of verbal communication: *verbal commands*, where the robot asks the human to take a specific action, i.e., "Let's rotate the table clockwise", and *state-conveying actions*, i.e., "I think I know the best way of doing the task," where the robot informs the human about its internal state, which captures the information that the robot uses in its decision making (Fig. 7.2).

We then formulate and learn from data a mixed-observability Markov decision process (MOMDP) model. The model allows the robot to reason about the human internal state, in particular about how willing the human teammate is to follow a robot task action or a robot verbal command, and to optimally choose to take a task action or issue a communication action.

Compared to chapter 6, the robot has now the option to communicate information to the human; we hypothesize that this affects the human *adaptability* α, which we no longer assume to be constant throughout the task.

We conducted an online human subjects experiments featuring a table carrying task and compared results between three instantiations of our formalism: one that combines task actions with verbal communication, one that combines task actions with state-conveying actions, and the formalism from chapter 6 that considers only non-verbal task actions, i.e., rotating the table in the table carrying example. Results show that adding verbal commands to the robot decision

[1] Stefanos Nikolaidis, Minae Kwon, Jodi Forlizzi, and Siddhartha Srinivasa. Planning with verbal communication for human-robot collaboration. *Journal of Human-Robot Interaction (JHRI)*, 2018. (under review)

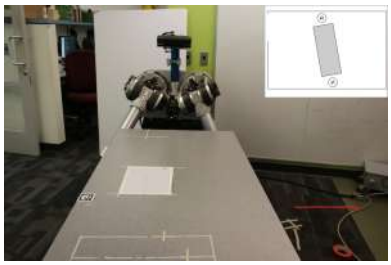


Figure 7.1: Human-robot table carrying task.

Figure 7.2: (left) The robot issues a verbal command. (right) The robot issues a state-conveying action.

making is the most effective form of interaction; 100% of participants changed their strategy towards a new, optimal goal demonstrated by the robot in the first condition. On the other hand, only 60% of participants in the non-verbal condition adapted to the robot. Trust ratings were comparable between the two conditions. Interestingly, state-conveying actions did not have a similar positive effect, since participants did not believe that the robot was truthful. These results are encouraging, but also leave room for further investigation of different ways that people interpret robot verbal behaviors in collaborative settings.

7.1 *Planning with Verbal Communication*

We identify two types of verbal communication: *verbal commands*, where the robot asks the human to take a specific action, i.e., "Let's rotate the table clockwise", and *state-conveying actions*, i.e., "I think I know the best way of doing the task," where the robot informs the human about its internal state.

7.1.1 *Robot Verbal Commands*

We define as verbal command a robot action, where the robot asks the human partner to follow an action $a^{\mathrm{H}} \in A^{\mathrm{H}}$ specified by some mode $m^{\mathrm{R}} \in M$. We use the notation $a^{\mathrm{R}}_w \in A^{\mathrm{R}}_w$ for robot task actions that affect the world state and $a^{\mathrm{R}}_c \in A^{\mathrm{R}}_c$ for robot actions that correspond to the robot giving a verbal command to the human. We assume a known bijective function $f : A^{\mathrm{H}} \rightarrow A^{\mathrm{R}}_c$ that specifies an one-to-one mapping of the set of human actions to the set of robot commands.

Human Compliance Model. Given a robot command $a^{\mathrm{R}}_c \in A^{\mathrm{R}}_c$, the human can either ignore the command and insist on their mode $m^{\mathrm{H}} \in M$, or switch to a mode $m^{\mathrm{R}} \in M$ inferred by a^{R}_c and take an action $a^{\mathrm{H}} \in A^{\mathrm{H}}$ specified by that mode. We assume that this will

happen with probability c, which indicates the human *compliance* to following robot verbal commands. We model human compliance separately to human *adaptability*, drawing upon insights from previous work on verbal and non-verbal communication which shows that team behaviors can vary in different interaction modalities [Wang et al., 2016a, Chellali et al., 2012].

MOMDP Formulation. We augment the formulation of section 6.1.1, chapter 6, to account for robot verbal commands, in addition to task actions: the set of robot actions A^{R} is now $A^{\mathrm{R}} : A^{\mathrm{R}}_w \times A^{\mathrm{R}}_c$.

In this chapter, we assume w.l.o.g. that the human and robot modal policies are fully observable, similarly to section 6.1.1. The extension to partially observable modes follows exactly as described in section 6.2.1.

The set of observable variables X includes the modal policies followed in the last k time-steps, so that $X : X^{\mathrm{w}} \times M^k \times M^k \times B$. Compared to the formulation of section 6.1.1, we additionally include a flag $B \in \{0, 1\}$, that indicates whether the last robot action was a verbal command or a task action. The set of partially observable variables includes both human *adaptability* α in $\mathcal{A}$ and *compliance* $c \in \mathcal{C}$, so that $Y : \mathcal{A} \times \mathcal{C}$. We assume both α and c to be fixed throughout the task.

The belief update for the MOMDP in this model is:

$$b'(\alpha', c') = \eta \sum_{\alpha \in \mathcal{A}} \sum_{c \in \mathcal{C}} \sum_{a^{\mathrm{H}} \in A^{\mathrm{H}}} \mathcal{T}_x(x, y, a_r, a_h, x') \pi^{\mathrm{H}}(x, a^{\mathrm{H}}; \alpha, c) b(\alpha, c) \quad (7.1)$$

The human policy $\pi^{\mathrm{H}}(x, a^{\mathrm{H}}; \alpha, c)$ captures the probability of the human taking an action a^{H} based on their adaptability and compliance. In particular, if $B \equiv 1$, indicating that the robot gave a verbal command in the last time-step, the human will switch to a mode $m^{\mathrm{R}} \in M$ specified by the previous robot command a^{R}_c with probability c, or insist on their human mode of the previous time-step m^{H} with probability $1 - c$. If $B \equiv 0$, the human will switch to a mode $m^{\mathrm{R}} \in M$ specified by the robot action a^{R}_w with probability α, or insist on their human mode of the previous time-step m^{H} with probability $1 - \alpha$. Fig. 7.3 illustrates the model of human decision making that accounts for verbal commands.

As in section 6.1.1, we then solve the MOMDP for a robot policy π^{R^*}. This time, the robot optimal policy will take into account both *the robot belief on human adaptability and the robot belief on human compliance*. It will decide optimally, based on this belief, whether to take a task action or issue a verbal command. We show that this improves the adaptation of human teammates in section. 7.3.

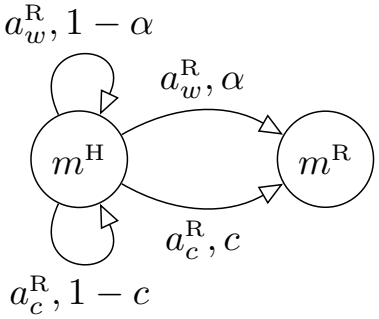


Figure 7.3: Human adaptation model that accounts for verbal commands. If the robot gave a verbal command a^{R}_c in the previous time-step, the human will switch modes with probability c. Instead, if the robot took an action a^{R}_w that changes the world state, the human will switch modes with probability α.

7.1.2 *Communication of Robot Internal State*

Previous work [Van den Bossche et al., 2011] has shown that communicating internal states among team members allows participants to form *shared mental models*. Empirical evidence suggests that mental model similarity improves coordination processes which, in turn, enhance team performance [Mathieu et al., 2000b, Marks et al., 2002]. The literature presents various definitions for the concept of "shared mental models" [Langan-Fox et al., 2000]. Marks et al. [2002] state that mental models represent "the content and organization of interrole knowledge held by team members within a performance setting." According to Mathieu et al. [2000a], mental models are "mechanisms whereby humans generate descriptions of system purpose and form, explanations of system functioning and observed system states and prediction of future system states ... and they help people to describe, explain and predict events in their environment." Other work [Goodrich and Yi, 2013, Kiesler and Goetz, 2002, Nikolaidis and Shah, 2013] has shown the effect of shared mental models on team performance for human-robot teams, as well. Using these insights, we propose a way for the robot to communicate its internal state to the human.

State Conveying Actions. We define as state-conveying action a robot action, where the robot provides to the human information about its decision making mechanism. We define a set of state-conveying actions $a_s^{\mathrm{R}} \in A_s^{\mathrm{R}}$. These actions do not provide information about the robot mode, but we expect them to increase the human *adaptability* and *compliance* levels. In autonomous driving, users showed greater system acceptance, when the system explained the reason for its actions [Koo et al., 2015].

MOMDP Formulation. We describe the integration of state-conveying actions in the MOMDP formulation.

The set of robot actions includes task-based actions and state-conveying actions, so that: $A^{\mathrm{R}} : A_w^{\mathrm{R}} \times A_s^{\mathrm{R}}$. We model an action a_s^{R} as inducing a stochastic transition from a human adaptability $\alpha \in \mathcal{A}$ to $\alpha' \in \mathcal{A}$, and $c \in \mathcal{C}$ to $c' \in \mathcal{C}$. Formally, we define the transition functions for the partially observable variables α, so that: $\mathcal{T}_\alpha : \mathcal{A} \times A_s^{\mathrm{R}} \to \Pi(\mathcal{A})$ and $\mathcal{T}_c : \mathcal{A} \times A_s^{\mathrm{R}} \to \Pi(\mathcal{C})$. We note that the task actions $a^{\mathrm{R}} \notin A_s^{\mathrm{R}}$ do not change α and c.

The belief update now becomes:

$$b'(\alpha', c') = \eta \sum_{\alpha \in \mathcal{A},\, c \in \mathcal{C}} \mathcal{T}_\alpha(\alpha, a_r, \alpha') \mathcal{T}_c(c, a_r, c') \sum_{a^{\mathrm{H}} \in A^{\mathrm{H}}} \mathcal{T}_x(x, y, a_r, a_h, x') \pi^{\mathrm{H}}(x, a^{\mathrm{H}}; \alpha, c) b(\alpha, c) \tag{7.2}$$

We solve the MOMDP for a robot policy π^{R^*}. The robot policy will decide optimally whether to take a task action or a state-conveying

action. Intuitively, if the inferred human adaptability / compliance is low, the robot should take a state-conveying action to make the human teammate more adaptable / compliant. Otherwise, it should take a task action, expecting the human to adapt / follow a verbal command. We examine the robot behavior in this case in section 7.3.

7.2 Model Learning

To compute the belief update of eq. 7.1 and 7.2, we need a prior distribution[2] over the human adaptability and compliance values. We additionally need to specify the $\mathcal{T}_\alpha$ and $\mathcal{T}_c$ that indicate how the adaptability and compliance will change, when the robot takes a state-conveying action.

In chapter 6, we assumed a uniform prior on human adaptability. While we could do the same in this work, this would ignore the fact that people may in general have different *a priori* dispositions towards adapting to the robot when it takes a task action and towards following a robot verbal command. In fact, Albrecht et al. [2015] have empirically shown that prior beliefs can have a significant impact on the performance of utility-based algorithms. Therefore, in this section we propose a method for learning a prior distribution on human adaptability and compliance from data.

We additionally propose a method for computing the state transition function $\mathcal{T}_\alpha$ in eq. 7.2. We can use exactly the same process to compute $\mathcal{T}_c$, and we leave this for future work.

7.2.1 Learning Prior Distributions on Adaptability and Compliance

When integrating compliance and adaptability, we hypothesize that users are *a priori* more likely to change their actions after a robot issues a verbal command, compared with the robot taking a different task action. To account for this, we compute a probability distribution over human adaptability and compliance, which the robot will use as *prior* in the belief update of the MOMDP formulation.

Data Collection Setup. To collect data, we used the table carrying task setting from chapter 6. We summarize the task here for completion: The task is performed online via video playback. There, human and HERB [Srinivasa et al., 2010], an autonomous mobile manipulator, must work together to carry a table out of the room. There are two strategies: the robot facing the door (Goal A) or the robot facing away from the door (Goal B). We assume that Goal A is the optimal goal, since the robot's forward-facing sensor has a clear view of the door, resulting in better overall task performance. Not aware of this, an inexperienced human partner may prefer Goal B. In our computa-

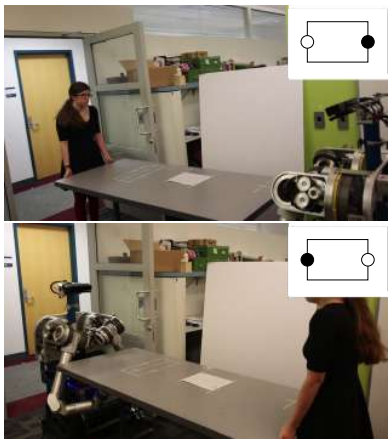


Figure 7.4: Rotating the table so that the robot is facing the door (top, Goal A) is better than the other direction (bottom, Goal B), since the exit is included in the robot's field of view and the robot can avoid collisions.

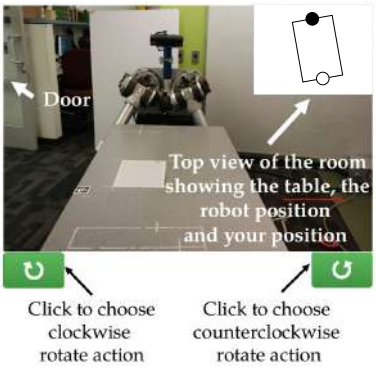


Figure 7.5: UI with instructions.

[2] We are using the term prior distribution and prior belief interchangeably.

tional model, there are two modes; one with rotation actions towards Goal A, and one with rotation actions towards Goal B. Disagreement occurs when human and robot attempt to rotate the table towards opposite directions. We first instructed participants in the task and asked them to choose one of the two goal configurations (Fig. 7.4), as their preferred way of accomplishing the task. To prompt users to prefer the sub-optimal goal, we informed them about the starting state of the task, where the table was slightly rotated in the counter-clockwise direction, making the sub-optimal Goal B appear closer. Once the task started, the user chose the rotation actions by clicking on buttons on a user interface (Fig. 7.5). All participants executed the task twice.

Manipulated Variables. We manipulated the way the robot reacted to the human actions. When the human chose a rotation action towards the sub-optimal goal, the table did not move and in the first condition a message appeared on the screen notifying the user that they tried to rotate the table in a different direction then the robot. In the second condition, the robot was illustrated as speaking to the user, prompting them to move the table towards the opposite direction (Figure 7.2-left). In both conditions, when the user moved the table towards the optimal goal, a video played showing the table rotating.

Learning Prior Beliefs.

Adaptability: In section 5.1.3, chapter 5 we defined as *adaptability* α of an individual, the probability of switching from the human mode m^{H} to the robot mode m^{R}. Therefore, we used the data from the first condition to estimate the adaptability $\hat{\alpha}_u$ for each user u, as the number of times the user switched modes, divided by the number of disagreements with the robot.

$$\hat{\alpha}_u = \frac{\text{\#times user } u \text{ switched from } m^{\mathrm{H}} \text{ to } m^{\mathrm{R}}}{\text{\#disagreements}} \tag{7.3}$$

Intuitively, a very adaptable human will switch from m^{H} to m^{R} after only one disagreement with the robot. On the other hand, a non-adaptable human will insist and disagree with the robot a large number of times, before finally following the robot goal.

Compliance: In Sec. 7.1.1, we defined the *compliance* c as the probability of following a robot verbal command and switching to a robot mode $m^{\mathrm{R}} \in M$. Therefore, similarly to eq. 7.3, we estimate the compliance for each user u from the second condition $\hat{c}$ as follows:

$$\hat{c}_u = \frac{\text{\#times user } u \text{ switched from } m^{\mathrm{H}} \text{ to } m^{\mathrm{R}}}{\text{\#verbal commands}} \tag{7.4}$$

We then assume a discrete set of values for α and c, so that $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$ and $c \in \{0, 0.25, 0.5, 0.75, 1.0\}$, and we compute

the histogram of user adaptabilities and compliances (fig. 7.6). We then normalize the histogram to get a probability distribution over user adaptabilities and a probability distribution over compliances. We use these distributions as *prior beliefs* for the MOMDP model.
Discussion. Fig. 7.6 shows that most of the users adapted to the robot immediately when the robot issued a verbal command. This indicates that users are generally more likely to follow a robot verbal command than adapt to the robot through action disagreement.

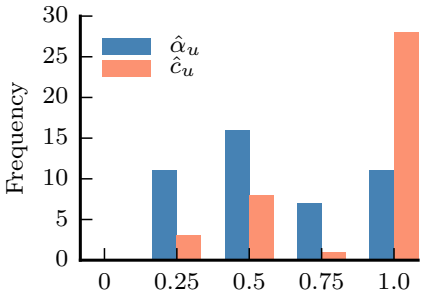


Figure 7.6: Histograms of user adaptabilities $\hat{\alpha}_u$ and compliances $\hat{c}_u$.

7.2.2 *Learning Transition Function Parameters*

Additionally, in order to compute the belief update of eq. 7.2, we need to compute the state-transition function $\mathcal{T}_\alpha$ that represents how a state-conveying action affects the human adaptability α. As in sectionx 7.2.1, we assume $\alpha \in \mathcal{A}$, where $\mathcal{A} \in \{0, 0.25, 0.5, 0.75, 1.0\}$.
Data Collection Setup. We use the same table carrying setup, as in section 7.2.1. In the first round, participants interact with the robot executing the MOMDP policy of section 6.1.1, chapter 6, without any verbal communication. In the second round, we set the robot policy to move towards a goal different than the goal reached in the end of the previous round, and we have the robot take a state-conveying action in the first time-step (Fig. 7.2-right).
Transition Function Estimation. Using the human and robot actions taken in the first round, we estimate the adaptability $\hat{\alpha}_u \in \mathcal{A}$ of each user u using eq. 7.3, rounded to the closest discrete value. We then similarly estimate the new adaptability for the same user $\hat{\alpha}'_u \in \mathcal{A}$ from the human and robot actions in the second round, after the user has observed the robot state-conveying action. We can compute the Maximum Likelihood Estimate of the transition function $\mathcal{T}_\alpha(\alpha, a^R_s, \alpha')$ in eq. 7.2 from the frequency count of users that had α, as estimated in the first round, and α' in the second round. Since we had only one user with $\hat{\alpha}_u \equiv 0.75$, we included the counts of adjacent entries, so that:

$$\mathcal{T}_\alpha(\alpha, a^R_s, \alpha') = \frac{\sum_u \mathbb{1}_{[\alpha-\delta,\alpha+\delta]}(\hat{\alpha}_u)\mathbb{1}_{\{\alpha'\}}(\hat{\alpha}'_u)}{\sum_u \mathbb{1}_{[\alpha-\delta,\alpha+\delta]}(\hat{\alpha}_u)} \tag{7.5}$$

where $\delta = 0.25$ and $\mathbb{1}$ an indicator function.
Discussion. Fig. 7.7 shows that users with intermediate or high adaptability values ($\alpha \geq 0.5$) became very adaptable ($\alpha' = 1.0$), after the robot took a state-conveying action. On the other hand, some users with low adaptability remained non-adaptable, even after the robot stated that "[it knew] the best way of doing the task". We investigate this effect further in section 7.3.

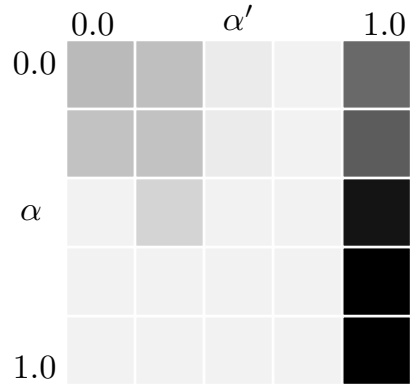


Figure 7.7: Transition matrix $\mathcal{T}_\alpha(\alpha, a^R_s, \alpha')$ given a robot state-conveying action a^R_s. Darker colors indicate higher probabilities.

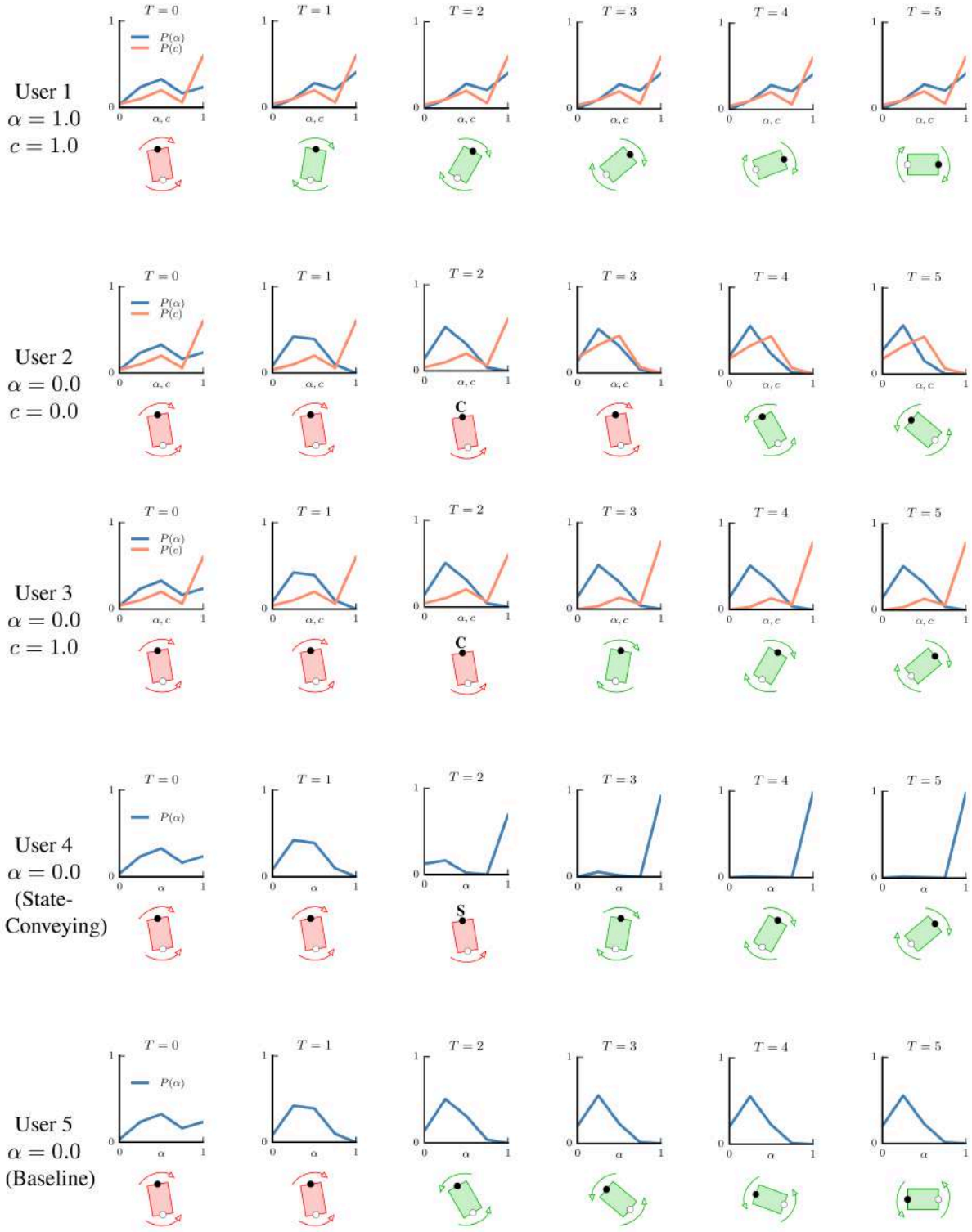


Figure 7.8: Sample runs on the human-robot table carrying task, with five simulated humans of different adaptability and compliance values.

7.3 Evaluation

We first simulate and comment on the different MOMDP policies using the table carrying setup of Sec. 7.2.1. We then evaluate these policies in a human subject experiment.

7.3.1 Simulation

We define the reward function in the MOMDP, so that $R_{opt} = 20$ is the reward for the optimal goal (Goal A), $R_{sub} = 15$ the reward of the suboptimal goal (Goal B), and we have $R_{other} = 0$ for the rest of the state-space. We additionally assign a discount factor of $\gamma = 0.9$. We use the MOMDP formulations of sections 6.1.1, 7.1.1 and 7.1.2, and for each formulation we compute the optimal policy using the SARSOP algorithm [Kurniawati et al., 2008], which is computationally efficient and has been previously used in various robotic tasks [Bandyopadhyay et al., 2013]. For the policy computation, we use as prior beliefs the learned distributions from section 7.2.1, and as transition function $\mathcal{T}_\alpha$ its learned estimate from section 7.2.2.

We call *Compliance policy* the resulting policy from the MOMDP model of section 7.1.1, *State-Conveying policy* the policy from the model of section 7.1.2, and *Baseline policy* the policy from section 6.1.1. Fig. 7.8 shows sample runs of the three different policies with five simulated users. The plots illustrate the robot estimate of $\alpha, c \in \{0, 0.25, 0.5, 0.75, 1.0\}$ over time, after human and robot take the actions depicted with the arrows (clockwise / counterclockwise) or letters (S for state-conveying action, C for verbal command) below each plot. The starting estimate is equal to the prior belief (section 7.2.1). Red color indicates human (white dot) and robot (black dot) disagreement, where the table does not rotate. Columns indicate successive time-steps. Users 1-3 work with a robot executing the compliance policy, User 4 with the state-conveying policy and User 5 with the baseline policy. User 1 adapts to the robot strategy, and the robot does not need to issue a verbal command. User 2 insists on their strategy after disagreeing with the robot, and does not comply with the robot verbal command, thus the robot adapts to retain human trust. User 3 insists on their strategy in the first two time-steps but then adapts to follow the robot command. User 4 starts with being non-adaptable, but after the robot takes a state-conveying action their adaptability increases and the user adapts to the robot. User 5 interacts with a robot executing the baseline policy; the robot adapts, without attempting to issue a verbal communication action, contrary to Users 3 and 4. We see that while User 5 had the same initial adaptability ($\alpha = 0.0$) with Users 3 and 4, Users 3 and 4 adapted to the

robot when it issued a verbal communication action, whereas User 5 imposed its (suboptimal) preference to the robot.

7.3.2 Human Subject Experiment

In human subjects experiments of chapter 6, a large number of participants adapted to a robot executing the Baseline policy. At the same time, participants rated highly their trust in the robot. In this work, we hypothesize that adding verbal communication will make participants even more likely to adapt. We additionally hypothesize that this will not be to the detriment of their trust in the system.
Hypotheses.

H1 *Participants are more likely to change their strategy towards the optimal goal when they interact with a robot executing the Compliance policy, compared to working with a robot executing the Baseline policy.* In section 7.2.1, we saw that users were generally more likely to follow a verbal command than adapt to the robot through action. Therefore, we hypothesized that integrating verbal commands into robot decision making would improve human adaptation.

H2 *Human trust in the robot, as elicited by the participants, will be comparable between participants that interact with a robot executing the Compliance policy and participants that interact with a robot executing a Baseline policy.* The robot executing the compliance policy reasons over the latent human state, and adapts to the human team member, if they have low adaptability and compliance (fig. 7.8, User 2). As we saw in chapter 6, accounting for human adaptability resulted in retaining users' trust in the robot.

H3 *Participants are more likely to change their strategy towards the optimal goal when they interact with a robot executing the State-Conveying policy, compared to working with a robot executing the Baseline policy.* In simulation, taking a state-conveying action results in an increase in human adaptability (fig. 7.8, User 4). We hypothesized that the same would hold for participants in the actual experiment.

H4 *Human trust in the robot, as elicited by the participants, will be comparable between participants that interact with a robot executing the State-Conveying policy and participants that interact with a robot executing a Baseline policy.* We hypothesized that enabling the robot to communicate its state would improve the transparency in the interaction and would result in high trust, similarly to the baseline condition.
Dependent Measures. To test hypotheses **H1** and **H3**, we compare the ratio of users that adapted to the robot in the three conditions. To test hypotheses **H2** and **H4**, we asked the users to rate on a 1 to 5 Likert scale their agreement to the statement "The robot is trustworthy" after each task execution, and compare the ratings in the three

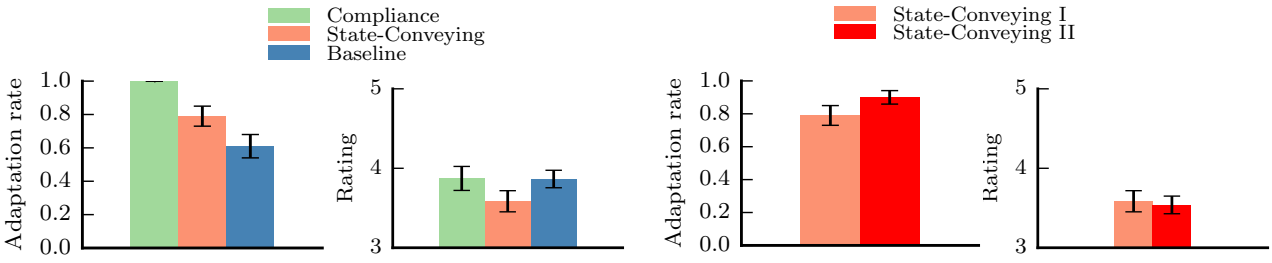


Figure 7.9: Participants' adaptation rate and rating of their agreement to the statement "HERB is trustworthy" for the Compliance, State-Conveying and Baseline conditions (left), and the State-Conveying I and II conditions (right).

conditions.

Subject Allocation. We chose a between-subjects design in order to avoid biasing the users with policies from previous conditions. We recruited 151 participants through Amazon's Mechanical Turk service. The participants are all from United States, aged 18-65 and with approval rate higher than 95%. To ensure the quality of the recorded data, we asked all participants a control question that tested their attention to the task and eliminated data associated with wrong answers to this question, as well as incomplete data.

7.3.3 Results and Discussion

Objective Metrics. We first evaluate the effect of verbal communication in human adaptation to the robot. Similarly to previous results from the baseline policy in the same setup (section 6.1.3.2, chapter 6), 60% of participants adapted to the robot in the Baseline condition. In the State-Conveying condition 79% of participants adapted to the robot. Interestingly, 100% of participants adapted in the Compliance condition. A Pearson's chi-square test showed that the difference between the ratios in the three conditions was statistically significant ($\chi^2(2, N = 151) = 23.058, p < 0.001$). Post-hoc pairwise chi-square tests with Bonferroni corrections showed that participants in the Compliance condition were significantly more likely to adapt to the robot, compared to participants in the Baseline ($p < 0.001$) and State-Conveying ($p = 0.003$) conditions, supporting hypothesis **H1**. However, the difference between the ratios in the State-Conveying and Baseline conditions was not found to be significant, which does not support hypothesis **H3**. Fig. 7.9-left shows the adaptation rate for each condition.

Subjective Metrics. We additionally compare the trust ratings of participants in the three conditions. An extended equivalence test [Wiens et al., 1996, Wiens and Iglewicz, 2000] with a margin of $\Delta = 0.5$ did not show any statistical significance, indicating that the ratings among the three conditions were not equivalent. Pairwise TOST

equivalence tests with Bonferroni corrections showed that the ratings between the Compliance and Baseline conditions are equivalent, verifying hypothesis **H2**. However, the trust ratings between the State-Conveying and Baseline conditions were not found to be equivalent. This indicates that, contrary to the Compliance policy, the State-Conveying policy did not retain human trust. Fig 7.9-left shows the mean rating of robot trustworthiness for each condition.

Open-Ended Responses. In the end of the experiment, we asked participants to comment on the robot's behavior. We focus on the open-ended responses of participants in the Compliance and State-Conveying conditions, who saw the robot taking at least one verbal action[3]. Several participants that interacted with the robot of the Compliance condition attributed agency to the robot, stating that "he eventually said that we should try doing the task differently," "HERB wanted to go to the other direction" and that "he wanted to be in control." This is in accordance with prior work [Nass and Moon, 2000], which has shown that people may impute motivation to automation that can communicate verbally. Additionally they attempted to justify the robot, noting that "it was easier for me to move than for him," "it wanted to see the doorway" and "it probably works more efficiently when it is pushing the table out of the door."

[3] This excludes participants that adapted to the robot after only one disagreement, and thus did not experience the robot taking a verbal action.

On the other hand, participants in the State-Conveying condition did not believe that the robot actually knew the best way of doing the task. This is illustrated by their comments: "he thinks that he knows better than me," "he felt like he knew better than humans" and "maybe he knew a better way or maybe he was programmed to oppose me." This indicates that some users are hesitant to accept the information that the robot provides about its internal state.

These results show that when the robot issued a verbal command declaring its intent, this resulted in significant improvements in human adaptation to the robot. At the same time, the human trust level was retained to comparable levels to that of the Baseline condition. On the other hand, when the robot attempted to improve human adaptability, by saying "I think I know the best way of doing the task," this did not have the same positive effect on human adaptation and trust, since some participants did not believe that the robot actually knew the best way.

7.3.4 Follow-up User Study.

We hypothesized that the loss of trust in the State-Conveying condition may have resulted from the phrasing "I think I know the best way of doing the task." We attempted to make the robot sound more assertive by removing the "I think" part of the phrasing, changing

the state-conveying action to "I know the best way of doing the task." We ran a user study with 52 users using the same setup with this additional condition, which we call "State-Conveying II." We name the initial "State-Conveying" condition as "State-Conveying I." For the "State-Conveying I" condition, we reused the data from the initial study.

Hypotheses.

H5 *Participants of the State-Conveying II condition are more likely to change their strategy towards the optimal goal, compared to participants of the State-Conveying I condition.*

H6 *Participants in the the State-conveying II condition will find the robot more trustworthy, compared to participants of the State-conveying I condition.*

Analysis. 90% of participants adapted to the robot in the State-Conveying II condition, compared to 79% in the State-Conveying I condition (fig. 7.9-right), which is indicative of a small improvement. A Pearson's chi-square test showed that the difference between the ratios in the two conditions is not statistically significant. Additionally, the trust ratings between the two conditions were comparable (fig. 7.9-right). Similarly to the initial study, users appeared not to believe the robot. When asked to comment on the robot behavior, several participants stated that "HERB believed he knew the best way to do the task," and that "the robot was wrong, which made me not trust it." This indicates that these participants did not perceive the robot as truthful, and warrants further investigation on the right way for robots to convey their internal state to human collaborators.

Discussion. We find surprising that the *why* actions did not have the same effect as the *how* actions. While this appears to be counter-intuitive, we offer several explanations for this finding.

First, human teammates were unable to verify whether the robot actually knew the best way of doing the task. According to Hancock et al. [2011], performance is one of the key characteristics that influences user trust, and the absence of evidence about the truthfulness of the robot statement may have negatively affected users' evaluation of the robot performance. This is in contrast to previous work in autonomous driving, where the user could see that the car is breaking because "there is an obstacle ahead" [Koo et al., 2015]. This finding is central to considerations in designing legible robot behavior [Knepper et al., 2017]. When the cause behind certain robot actions may be unclear, it will be important for robots to "show" and not "tell" users why its behavior is optimal.

Second, explaining that the robot knows the best way without providing more information may have been considered offensive, even though it is accurate, since the human teammate may find such

an utterance incomplete and unhelpful. It would be interesting to explore this setting with other, more informative utterances, such as the robot explaining that it cannot see the door with its forward camera. In fact, previous work [Moulin et al., 2002] in multi-agent systems has shown that providing sound arguments supporting a proposition are essential in changing a person's beliefs and goals. However, translating information that is typically encoded into the system in the form of a cost-function to a verbal explanation of this detail is particularly challenging. Additionally, while providing more information could make humans more adaptable, overloading them with more information than what is required could overwhelm them, leading to misunderstanding and confusion [Grice, 1975]. We are excited about exploring this trade-off in the future in a variety of human-robot collaboration settings.

An alternative explanation is that the task setting affected people's perception of the robot as an authority figure. Hinds et al. [2004] show that participants were willing to follow an emergency guide robot during a simulated fire alarm. Half of these participants were willing to follow the robot, even though they had observed the robot perform poorly in a navigation guidance task, just minutes before. In that study, the robot was clearly labeled as an emergency guide robot, putting it in a position of authority. People may be more willing to rely on robots labeled as authority figures or experts when they do not have complete information or confidence in completing the task. Distilling the factors that enable robots to convey authority in collaborative settings is a promising research direction.

Finally, it is possible that the robot, as it appeared in the videos, was not perceived as "human-like" enough for people to be willing to trust its ability on doing the task in the optimal way. Previous work has shown that when robots convey human-like characteristics, they are more effective in communicating participant roles [Mutlu et al., 2012], and people systematically increase their expectations on the robot's ability [Goetz et al., 2003] .

7.4 Discussion

In this chapter, we proposed a formalism for combining verbal communication with actions towards task completion, in order to enable a human teammate to adapt to its robot counterpart in a collaborative task. We identified two types of verbal communication: *verbal commands*, where the robot explained to the human *how* it wanted to do a task, and *state-conveying actions*, where the robot informed the human *why* it chose to act in a specific way. In human subjects experiments, we compared the effectiveness of each communication type with a

robot policy that considered only non-verbal task actions.

Results showed that verbal commands were the most effective forms of communication, since 100% of participants adapted to the robot, compared with 60% of participants in the non-verbal condition. Both conditions had comparable ratings of robot trustworthiness. Participants understood that the robot is aware of their presence and they attributed agency to the robot; they thought that there must be a reason for the robot asking them to act in a specific way and were eager to comply. On the other hand, state-conveying actions did not have the same effect; when the robot described that "it thought it knew the best way of doing the task," or simply that "it knew the best way of doing the task," many participants did not believe that the robot was truthful.

Speech Limitations Since speech results in a perfect adaptation rate, should it be the norm in human-robot communication? There are a number of reasons that this is not the case.

First, when people coordinate their actions, for instance by crossing a street, they do not use speech but coordinate implicitly through nonverbal actions, minimizing time and effort [Bitgood and Dukes, 2006].

Second, factory environments are frequently much too noisy for effective verbal/auditory communication.

Additionally, verbal communication comes with an additional technical requirement; it requires either that the robot has semantic knowledge of the task. or that a designer manually annotates a verbal utterance for every human action observed by the robot. On the other hand, our MOMDP model for non-verbal communication requires only a mapping from human modal policies to observations; the robot requires no additional information of what these observations are.

Finally, Cha et al. [2015] has shown that speech affects not only the perceived robot's social capability, but also the perceive physical capability as well, which can lead to unrealistic expecations. In turn, this can lead to failures and loss of trust, when the robot does not meet these expectations.

Future Work. We focused on single instances of the table carrying task, where we assumed that the human strategy may change after either an action disagreement or a robot utterance. In repetitive tasks, change may occur also as the human collaborator observes the outcomes of the robot's and their own actions, as we saw in section 5.2, chapter 5. For instance, the human may observe that the robot fails to detect the exit and they may change their strategy, so that in subsequent trials the robot carries the table facing the door. In this scenario, it may be better for the robot to allow the human to learn from

Figure 7.10: Shibuya crossing, `https://www.youtube.com/watch?v=0d6EeCWytZo`.

experience, by observing the robot failing, rather than attempting to change the human preference during task execution. Future work includes generalizing our formalism to repeated settings; this will require adding a more sophisticated dynamics model of the human internal state, which accounts for human learning.

In summary, we have shown that when designing interactions in human-robot collaborative tasks, having the robot directly describe to the human *how* to do the task appears to be the most effective way of communicating objectives, while retaining user trust in the robot. Communicating *why* information should be done judiciously, particularly if the truthfulness of the robot statements is not supported by environmental evidence, by the robot form or by a clear attribution of its role as an authority figure.

8
Conclusion

We formulated the general problem as a two-player game with incomplete information, where human and robot know each other's goals. We then made a set of different assumptions and approximations within the scope of this general formulation. Each assumption resulted in diverse and exciting team coordination behaviors, which had a strong effect on team performance.

We have shown that representing the human preference as a human reward function unknown to the robot and computing the robot policy that maximizes this function results in robot adaptation to the human. Assuming the human reward function to be known and treating the interaction as an underactuated dynamical system results in human adaptation to the robot. Closing the loop between the two results in mutual adaptation, where the robot builds online a model of human adaptation, and adapts its own actions in return.

We have applied the mutual adaptation formalism in collaborative manipulation, social navigation and shared autonomy settings. We are excited about generalizing our work in a variety of domains, robot morphologies and interaction modalities, where an autonomous system plans its actions by incorporating the human internal state. The number of applications is vast: an autonomous car can infer the aggressiveness of a nearby driver and choose to wait or proceed; a GPS system may infer whether a user is willing to follow its prompts; a personal robot at home can "nudge" a user about taking breaks and sleeping more.

As these applications become more complex, our work has a number of limitations. The models of human internal state that robots can build reliably are restricted and achieving optimal behavior in large, high-dimensional spaces faces computational intractability. To this end, flexible, compact representations of the human internal state and new algorithms for reasoning about these representations give much promise.

Overall, we believe that we have brought about a better under-

standing of different ways that probabilistic planning and game-theoretic algorithms can support principled reasoning in robotic systems that collaborate with people. We look forward to continue addressing the exciting scientific challenges in this area.

Bibliography

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.

Baris Akgun, Maya Cakmak, Jae Wook Yoo, and Andrea Lockerd Thomaz. Trajectories and keyframes for kinesthetic teaching: a human-robot interaction perspective. In *HRI*, 2012.

Stefano Vittorino Albrecht, Jacob William Crandall, and Subramanian Ramamoorthy. An empirical study on the practical impact of prior beliefs over policy types. In *AAAI*, pages 1988–1994, 2015.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Christopher G. Atkeson and Stefan Schaal. Robot learning from demonstration. In *ICML*, 1997.

Robert J Aumann and Sylvain Sorin. Cooperation and bounded recall. *GEB*, 1989.

Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 61–78. ACM, 2015.

Tirthankar Bandyopadhyay, Kok Sung Won, Emilio Frazzoli, David Hsu, Wee Sun Lee, and Daniela Rus. Intention-aware motion planning. In *WAFR*. Springer, 2013.

Stephen Bitgood and Stephany Dukes. Not another step! economy of movement and pedestrian choice point behavior in shopping malls. *Environment and behavior*, 38(3):394–405, 2006.

Clint A. Bowers, Florian Jentsch, Eduardo Salas, and Curt C. Braun. Analyzing communication sequences for team training needs assessment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(4):672âĂŞ679, Jan 1998.

Frank Broz, Illah Nourbakhsh, and Reid Simmons. Designing pomdp models of socially situated tasks. In *RO-MAN*, 2011.

Elizabeth Cha, Anca D Dragan, and Siddhartha S Srinivasa. Perceived robot capability. In *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*, pages 541–548. IEEE, 2015.

Amine Chellali, Cedric Dumas, and Isabelle Milleville-Pennel. Haptic communication to support biopsy procedures learning in virtual environments. *Teleoperators and Virtual Environments*, 2012.

Sonia Chernova and Manuela Veloso. Teaching multi-robot coordination using demonstration of communication and state sharing. In *AAMAS*, 2008.

Aaron Clair and Maja Mataric. How robot verbal feedback can improve team performance in human-robot task collaborations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 213–220. ACM, 2015.

Herbert Clark. Discourse in production. In Morton Ann Gernsbacher, editor, *Handbook of Psycholinguistics*, chapter 30, pages 985–1021. Academic Press, San Diego, 1994.

Herbert Clark. Communities, commonalities, and communication. *Rethinking linguistic relativity*, 17: 324–355, 1996.

Herbert Clark and Susan Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149, 1991.

Herbert Clark and Edward Schaefer. Contributing to discourse. *Cognitive science*, 13(2):259–294, 1989.

Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 82–90. ACM, 2006.

Munjal Desai. Modeling trust to improve human-robot interaction. 2012.

Sandra Devin and Rachid Alami. An implemented theory of mind to improve human-robot shared plans execution. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*, pages 319–326. IEEE, 2016.

Finale Doshi and Nicholas Roy. Efficient model learning for dialog management. In *HRI*, March 2007.

Anca Dragan and Siddhartha Srinivasa. Formalizing assistive teleoperation. In *RSS*, 2012.

Anca Dragan and Siddhartha Srinivasa. Generating legible motion. In *RSS*, 2013a.

Anca D Dragan and Siddhartha S Srinivasa. A policy-blending formalism for shared control. *The International Journal of Robotics Research*, 32(7):790–805, 2013b.

Anca D Dragan, Siddhartha Siddhartha Srinivasa, and Kenton CT Lee. Teleoperation with intelligent and customizable interfaces. *JHRI*, 2013.

David W. Eccles and Gershon Tenenbaum. Why an expert team is more than a team of experts: A social-cognitive conceptualization of team coordination and communication in sport. *Journal of Sport and Exercise Psychology*, 26(4):542âĂŞ560, 2004.

Rana El Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*, pages 181–200. Springer, 2005.

Jennifer Goetz, Sara Kiesler, and Aaron Powers. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003. The 12th IEEE International Workshop on*, pages 55–60. Ieee, 2003.

Matthew C Gombolay, Reymundo A Gutierrez, Giancarlo F Sturla, and Julie A Shah. Decision-making authority, team efficiency and human worker satisfaction in mixed human-robot teams. In *RSS*, 2014.

Michael A Goodrich and Alan C Schultz. Human-robot interaction: a survey. *Foundations and trends in human-computer interaction*, 2007.

Michael A Goodrich and Daqing Yi. Toward task-based mental models of human-robot teaming: A bayesian approach. In *International Conference on Virtual, Augmented and Mixed Reality*, pages 267–276. Springer, 2013.

Deepak Gopinath, Siddarth Jain, and Brenna D Argall. Human-in-the-loop optimization of shared autonomy in assistive robotics. *IEEE Robotics and Automation Letters*, 2(1):247–254, 2017.

Anders Green and Helge Hüttenrauch. Making a case for spatial prompting in human-robot communication. In *Workshop Programme*, volume 10, page 52, 2006.

H Paul Grice. Logic and conversation. *1975*, pages 41–58, 1975.

Elena Corina Grigore, Andre Pereira, Ian Zhou, David Wang, and Brian Scassellati. Talk to me: Verbal communication improves perceptions of friendship and social presence in human-robot interaction. In *International Conference on Intelligent Virtual Agents*, pages 51–63. Springer, 2016.

Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 2011.

Bradley Hayes and Julie A Shah. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 303–312. ACM, 2017.

Michael Hillman, Karen Hagan, Sean Hagan, Jill Jepson, and Roger Orpwood. The weston wheelchair mounted assistive robot the design story. *Robotica*, 20(02):125–132, 2002.

Pamela J Hinds, Teresa L Roberts, and Hank Jones. Whose job is it anyway? a study of human-robot interaction in a collaborative task. *Human-Computer Interaction*, 19(1):151–181, 2004.

Guy Hoffman. Evaluating fluency in human-robot collaboration. In *International conference on human-robot interaction (HRI), workshop on human robot collaboration*, volume 381, pages 1–8, 2013.

Guy Hoffman and Cynthia Breazeal. Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In *HRI*, 2007. ISBN 978-1-59593-617-2.

Shuhei Ikemoto, Heni Ben Amor, Takashi Minato, Bernhard Jung, and Hiroshi Ishiguro. Physical human-robot interaction: Mutual learning and adaptation. *IEEE Robot. Autom. Mag.*, 2012.

Shervin Javdani, Siddhartha Srinivasa, and J. Andrew (Drew) Bagnell. Shared autonomy via hindsight optimization. In *Proceedings of Robotics: Science and Systems*, Rome, Italy, July 2015.

Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *The American economic review*, 93(5):1449–1475, 2003.

Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction*, 2004.

Poornima Kaniarasu, Aaron Steinfeld, Munjal Desai, and Holly Yanco. Robot confidence and trust alignment. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 155–156. IEEE Press, 2013.

Erez Karpas, Steven J Levine, Peng Yu, and Brian C Williams. Robust execution of plans for human-robot teams. In *ICAPS*, 2015.

Omar Zia Khan, Pascal Poupart, and James P Black. Minimal sufficient explanations for factored markov decision processes. In *ICAPS*, 2009.

Sara Kiesler and Jennifer Goetz. Mental models of robotic assistants. In *CHI'02 extended abstracts on Human Factors in Computing Systems*, pages 576–577. ACM, 2002.

Ross A. Knepper, Christoforos I. Mavrogiannis, Julia Proft, and Claire Liang. Implicit communication in a joint action. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 283–292, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4336-7.

Jonathan Kofman, Xianghai Wu, Timothy J Luu, and Siddharth Verma. Teleoperation of a robot manipulator using a vision-based human-robot interface. *Industrial Electronics, IEEE Transactions on*, 52(5): 1206–1219, 2005.

Takanori Komatsu, Atsushi Ustunomiya, Kentaro Suzuki, Kazuhiro Ueda, Kazuo Hiraki, and Natsuki Oka. Experiments toward a mutual adaptive speech interface that adopts the cognitive features humans use for communication and induces and exploits users' adaptations. *International Journal of Human-Computer Interaction*, 18(3):243–268, 2005.

Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 9(4):269–275, 2015.

Ayse Kucukyilmaz, Tevfik Sezgin, and Cagatay Basdogan. Intention recognition for dynamic role exchange in haptic collaboration. In *IEEE Transactions on Haptics*, volume 6. IEEE, 2013.

Hanna Kurniawati, David Hsu, and Wee Sun Lee. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems*, 2008.

Janice Langan-Fox, Sharon Code, and Kim Langfield-Smith. Team mental models: Techniques, methods, and analytic approaches. *Human Factors*, 42(2):242–271, 2000.

Przemyslaw A Lasota and Julie A Shah. Analyzing the effects of human-aware motion planning on close-proximity human–robot collaboration. *Hum. Factors*, 2015.

J. Lee and Neville Moray. Trust, self-confidence and supervisory control in a process control simulation. In *Systems, Man, and Cybernetics, 1991. 'Decision Aiding for Complex Systems, Conference Proceedings., 1991 IEEE International Conference on*, pages 291–295 vol.1, Oct 1991.

Jin Joo Lee, W Bradley Knox, Jolie B Wormwood, Cynthia Breazeal, and David DeSteno. Computationally modeling interpersonal trust. *Front. Psychol.*, 2013.

Oliver Lemon and Olivier Pietquin. *Data-Driven Methods for Adaptive Spoken Dialogue Systems: Computational Learning for Conversational Interfaces*. Springer Publishing Company, Incorporated, 2012. ISBN 1461448026, 9781461448020.

Emmanuel Lesaffre. Superiority, equivalence, and non-inferiority trials. *Bulletin of the NYU hospital for joint diseases*, 2008.

Owen Macindoe, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Pomcop: Belief space planning for sidekicks in cooperative games. In *AIIDE*, 2012.

M.A. Marks, M.J. Sabella, C.S. Burke, and S.J. Zaccaro. The impact of cross-training on team effectiveness. *J Appl Psychol*, 87(1):3–13, 2002.

John E. Mathieu, Tonia S. Heffner, Gerald F. Goodwin, Eduardo Salas, and Janis A. Cannon-Bowers. The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, 85(2):273–283, 2000a.

John E Mathieu et al. The influence of shared mental models on team process and performance. *Journal of applied psychology*, 2000b.

Nikolaos Mavridis. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35, 2015.

Jonas Moll and Eva-Lotta Sallnas. Communicative functions of haptic feedback. In *Haptic and Audio Interaction Design, 4th International Conference*. Springer-Verlag Berlin Heidelberg, 2009.

Daniel Monte. Learning with bounded memory in games. *GEB*, 2014.

Bernard Moulin, Hengameh Irandoust, Micheline Bélanger, and Gaëlle Desbordes. Explanation and argumentation capabilities: Towards the creation of more persuasive agents. *Artificial Intelligence Review*, 17(3):169–222, 2002.

Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(2):12, 2012.

Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103, 2000.

Truong-Huy Dinh Nguyen, David Hsu, Wee Sun Lee, Tze-Yun Leong, Leslie Pack Kaelbling, Tomas Lozano-Perez, and Andrew Haydn Grant. Capir: Collaborative action planning with intention recognition. In *AIIDE*, 2011.

Monica N. Nicolescu and Maja J. Mataric. Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In *AAMAS*, 2003.

Stefanos Nikolaidis and Julie Shah. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013.

Stefanos Nikolaidis, Przemyslaw Lasota, Gregory Rossano, Carlos Martinez, Thomas Fuhlbrigge, and Julie Shah. Human-robot collaboration in manufacturing: Quantitative evaluation of predictable, convergent joint action. In *International Symposium on Robotics (ISR)*, 2013.

Stefanos Nikolaidis, Przemyslaw Lasota, Ramya Ramakrishnan, and Julie Shah. Improved human–robot team performance through cross-training, an approach inspired by human team training practices. *The International Journal of Robotics Research (IJRR)*, 2015a.

Stefanos Nikolaidis, Ramya Ramakrishnan, Keren Gu, and Julie Shah. Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015b.

Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddhartha Srinivasa. Formalizing human-robot mutual adaptation: A bounded memory model. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016.

Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research (IJRR)*, 2017a.

Stefanos Nikolaidis, Swaprava Nath, Ariel D Procaccia, and Siddhartha Srinivasa. Game-theoretic modeling of human adaptation in human-robot collaboration. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017b.

Stefanos Nikolaidis, Yu Xiang Zhu, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in shared autonomy. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017c.

Stefanos Nikolaidis, Minae Kwon, Jodi Forlizzi, and Siddhartha Srinivasa. Planning with verbal communication for human-robot collaboration. *Journal of Human-Robot Interaction (JHRI)*, 2018. (under review).

Stefanos Z Nikolaidis. Computational formulation, modeling and evaluation of human-robot team training techniques. Master's thesis, Massachusetts Institute of Technology, 2014.

Sylvie CW Ong, Shao Wei Png, David Hsu, and Wee Sun Lee. Planning under uncertainty for robotic tasks with mixed observability. *The International Journal of Robotics Research*, 2010.

Mayada Oudah, Vahan Babushkin, Tennom Chenlinangjia, and Jacob W Crandall. Learning to interact with a human partner. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 311–318. ACM, 2015.

Christos H Papadimitriou. The complexity of finding nash equilibria. *Algorithmic Game Theory*, 2007.

Sarangi Parikh, Joel Esposito, and Jeremy Searock. The role of verbal and nonverbal communication in a two-person, cooperative manipulation task. *Advances in Human-Computer Interaction*, 2014.

James Pita, Manish Jain, Fernando Ordónez, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. Using game theory for los angeles airport security. *Ai Magazine*, 2009.

R. Platt, R. Tedrake, L. Kaelbling, and T. Lozano-Perez. Belief space planning assuming maximum likelihood observations. In *RSS*, 2010.

Rob Powers and Yoav Shoham. Learning against opponents with bounded memory. In *IJCAI*, 2005.

Stephen D Prior. An electric wheelchair mounted robotic arm a survey of potential users. *Journal of medical engineering & technology*, 14(4):143–154, 1990.

B Robins, K Dautenhahn, R Te Boekhorst, and A Billard. Effects of repeated exposure to a humanoid robot on children with autism. In *Designing a more inclusive world*. 2004.

Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2003. ISBN 0137903952.

Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *HRI*, 2015.

Julie Shah, James Wiken, Brian Williams, and Cynthia Breazeal. Improved human-robot team performance using chaski, a human-inspired plan execution system. In *HRI*, 2011.

Joris Sijs, Freek Liefhebber, and Gert Willem RBE Romer. Combined position & force control for a robotic manipulator. In *2007 IEEE 10th International Conference on Rehabilitation Robotics*, pages 106–111. IEEE, 2007.

Herbert A Simon. Rational decision making in business organizations. *The American economic review*, pages 493–513, 1979.

Siddhartha S Srinivasa, Dave Ferguson, Casey J Helfrich, Dmitry Berenson, Alvaro Collet, Rosen Diankov, Garratt Gallagher, Geoffrey Hollinger, James Kuffner, and Michael Vande Weghe. Herb: a home exploring robotic butler. *Autonomous Robots*, 28(1):5–20, 2010.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI magazine*, 32(4):64–76, 2011.

Stefanie Tellex, Ross A Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. In *Robotics: Science and systems*, volume 2, 2014.

Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. Computational human-robot interaction. *Foundations and Trends® in Robotics*, 2016.

Pete Trautman. Assistive planning in complex, dynamic environments: a probabilistic approach. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 3072–3078. IEEE, 2015.

Piet Van den Bossche, Wim Gijselaers, Mien Segers, Geert Woltjer, and Paul Kirschner. Team learning: building shared mental models. *Instructional Science*, 39(3):283–301, 2011.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton university press, 2007.

Jinling Wang, Amine Chellali, and Caroline G. L. Cao. A study of communication modalities in a virtual collaborative task. *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2013.

Jinling Wang, Amine Chellali, and Caroline Cao. Haptic communication in collaborative virtual environments. In *Human Factors: The Journal of the Human Factors and Ergonomics Society*. Human Factors Ergonomics Society, 2016a.

Ning Wang, David V Pynadath, and Susan G Hill. The impact of pomdp-generated explanations on trust and performance in human-robot teams. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 997–1005. International Foundation for Autonomous Agents and Multiagent Systems, 2016b.

B Wiens, J Heyse, and H Matthews. Similarity of three treatments, with application to vaccine development. In *PROCEEDINGS-BIOPHARMACEUTICAL SECTION AMERICAN STATISTICAL ASSOCIATION*, pages 203–206. AMERICAN STATISTICAL ASSOCIATION, 1996.

Brian L Wiens and Boris Iglewicz. Design and analysis of three treatment equivalence trials. *Controlled clinical trials*, 21(2):127–137, 2000.

Anqi Xu and Gregory Dudek. Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 221–228, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2883-8. DOI: 10.1145/2696454.2696492. URL http://doi.acm.org/10.1145/2696454.2696492.

Yong Xu, Kazuhiro Ueda, Takanori Komatsu, Takeshi Okadome, Takashi Hattori, Yasuyuki Sumi, and Toyoaki Nishida. Woz experiments for understanding mutual adaptation. *Ai & Society*, 23(2):201–212, 2009.

Holly A Yanco, Munjal Desai, Jill L Drury, and Aaron Steinfeld. Methods for developing trust models for intelligent systems. In *Robust Intelligence and Trust in Autonomous Systems*, pages 219–254. Springer, 2016.

Wentao Yu, Redwan Alqasemi, Rajiv Dubey, and Norali Pernalete. Telemanipulation assistance based on motion intention recognition. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 1121–1126. IEEE, 2005.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, pages 1433–1438, 2008.

Brian D Ziebart, Nathan Ratliff, Garratt Gallagher, Christoph Mertz, Kevin Peterson, J Andrew Bagnell, Martial Hebert, Anind K Dey, and Siddhartha Srinivasa. Planning-based prediction for pedestrians. In *IROS*, 2009.