Communication in Robot Learning : usage of the Combined Task and Social Channel for Action Learning

Manuel Bied

Sorbonne Université

École Doctorale SMAER (ED 391)

Institut des Systèmes Intelligents et de Robotique

Communication in Robot Learning: Usage of the Combined Task and Social Channel for Action Learning

Thèse de doctorat

Soutenue le 13. Janvier 2022 par

Christian Manuel BIED

Composition du Jury:

Ginevra CASTELLANO, Professeure *Rapporteuse*	Social Robotics Lab, Uppsala universitet, Sweden
Patrick HENAFF, Professeur *Rapporteur*	LORIA, Université de Lorraine, France
Francisco S. MELO, Professeur *Examinateur*	IST, Universidade de Lisboa, Portugal
Malika AUVRAY, Chargée de recherche *Examinatrice*	ISIR, Sorbonne Université, France
Mehdi KHAMASSI, Directeur de recherche *Examinateur*	ISIR, Sorbonne Université, France
Mohamed CHETOUANI, Professeur *Directeur de thèse*	ISIR, Sorbonne Université, France

Abstract

Creating artificial beings with human like capabilities is a long dream of humanity. While humanity already got closer to this goal, there is still a long way to go. A major problem preventing robots to be deployed in our day to day lives is the fact that robots can not be programmed for every situation they might encounter. This problem can be addressed by equipping robots with the capability to learn new tasks. The ability to learn is an important human social skill. Although what robots can and could accomplish focusing on manipulating their environment is already quite impressive, their social skills can be described as, if anything, rather basic. While there are already interactive robot learning approaches (e.g. reinforcement learning and learning from demonstration), usually these approaches do not consider explicit teaching. Furthermore, usually they exploit actions that either accomplish the task or serve a social purpose. However, human behavior allows for actions that combine task and social aspects in one action. Human-Robot Interaction (HRI) has not yet sufficiently addressed these combined actions.

In this thesis we focus on actions that combine task and social actions. We are interested in how humans use these type of actions in HRI settings, how teaching influences this behavior and how these actions can be used to to augment robot learning.

We conducted a user study investigating how humans behave when teaching how to solve a task to a robot in contrast to just solving the task. The study consisted of two experiments. In the first experiments the participants were first asked to solve a continuous maze task and to teach how to solve it to a robot afterwards. Furthermore, the participants could give negative demonstrations that demonstrated how the task should not be solved. In the second experiments the demonstrations collected in the first experiment were shown to new participants. The participants were asked how informative they perceive these demonstrations. The results show that significantly more negative than positive demonstrations were perceived as informative. Furthermore, significantly more demonstrations from the phase where the participants taught to a robot were perceived as informative than from the phase where the participants just solved the task.

We address the augmentation of robot learning with exploitation of task- and social channels by introducing a framework based on Reinforcement Learning (RL). In this framework we augment the reward from the environment with feedback how an observer might perceive the actions taken by the agent. We do this by proposing three different algorithms to model the observer perception as interactive RL scheme and compare with one non-interactive RL algorithm as baseline. In order to model the observer we vary the method how the observer estimates how likely the agent is going for the real goal. We evaluate our approach on five environments and calculate the legibility of the learned trajectories. Legibility is a scalar metric measuring how well goals can be inferred from actions. The results show that the legibility of the learned trajectories is significantly higher while integrating the feedback from the observer compared with a standard Q-Learning algorithm not using the observer feedback.

From these results we conclude that humans use combined actions in HRI settings to enrich the communication, but also perceive these actions as informative. Further, that combined actions can be learned with a RL framework by integrating reasoning about potential observers to enrich the actions with social aspects. While the research presented in thesis is limited to specific cases, it demonstrates the promising potential of combined actions in HRI settings.

Keywords: human-robot interaction, reinforcement learning, interactive robot learning, learning-from-demonstration, sensorimotor communication, legibility

Résumé (French Abstract)

Créer des êtres artificiels dotés de capacités humaines est un long rêve de l'humanité. Alors que l'humanité s'est déjà rapprochée de cet objectif, il reste encore un long chemin à parcourir. Un problème majeur empêchant le déploiement de robots dans notre vie de tous les jours est le fait que les robots ne peuvent pas être programmés pour chaque situation qu'ils pourraient rencontrer. Ce problème peut être résolu en dotant les robots de la capacité d'apprendre de nouvelles tâches. La capacité d'apprendre est une compétence sociale humaine importante. Bien que ce que les robots peuvent et pourraient accomplir en se concentrant sur la manipulation de leur environnement soit déjà assez impressionnant, leurs compétences sociales peuvent être décrites comme plutôt basiques. Bien qu'il existe déjà des approches interactives d'apprentissage par robot (par exemple, l'apprentissage par renforcement et l'apprentissage par démonstration), ces approches ne prennent généralement pas en compte l'enseignement explicite. De plus, ils exploitent généralement des actions qui accomplissent la tâche ou servent un objectif social. Cependant, le comportement humain permet des actions qui combinent tâches et aspects sociaux en une seule action. L'Interaction Homme-Robot (HRI) n'a pas encore suffisamment abordé ces actions combinées.

Dans cette thèse, nous nous concentrons sur les actions qui combinent tâches et actions sociales. Nous nous intéressons à la façon dont les humains utilisent ce type d'actions dans les environnements HRI, comment l'enseignement influence ce comportement et comment ces actions peuvent être utilisées pour augmenter l'apprentissage du robot.

Nous avons mené une étude utilisateur sur le comportement des humains lorsqu'ils enseignent à un robot comment résoudre une tâche, contrairement à la simple résolution de la tâche. L'étude consistait en deux expériences. Dans les premières expériences, les participants ont d'abord été invités à résoudre une tâche de labyrinthe continu et à enseigner ensuite comment la résoudre à un robot. De plus, les participants pouvaient faire des démonstrations négatives démontrant comment la tâche ne devrait pas être résolue. Dans la deuxième expérience, les démonstrations recueillies dans la première expérience ont été présentées à de nouveaux participants. On a demandé aux participants dans quelle mesure ils percevaient ces démonstrations informatives. Les résultats

montrent que significativement plus de démonstrations négatives que positives ont été perçues comme informatives. De plus, beaucoup plus de démonstrations de la phase où les participants ont enseigné à un robot ont été perçues comme informatives que de la phase où les participants viennent de résoudre la tâche.

Nous abordons l'augmentation de l'apprentissage des robots avec l'exploitation des tâches et des canaux sociaux en introduisant un cadre basé sur l'apprentissage par renforcement (RL). Dans ce cadre, nous augmentons la récompense de l'environnement avec une rétroaction sur la façon dont un observateur pourrait percevoir les actions entreprises par l'agent. Pour ce faire, nous proposons trois algorithmes différents pour modéliser la perception de l'observateur en tant que schéma RL interactif et comparer avec un algorithme RL non interactif comme référence. Afin de modéliser l'observateur, nous varions la méthode de la façon dont l'observateur estime la probabilité que l'agent se dirige vers l'objectif réel. Nous évaluons notre approche sur cinq environnements et calculons la lisibilité des trajectoires apprises. La lisibilité est une métrique scalaire mesurant dans quelle mesure les objectifs peuvent être déduits des actions. Les résultats montrent que la lisibilité des trajectoires apprises est significativement plus élevée tout en intégrant le feedback de l'observateur par rapport à un algorithme Q-Learning standard n'utilisant pas le feedback de l'observateur.

À partir de ces résultats, nous concluons que les humains utilisent des actions combinées dans les environnements HRI pour enrichir la communication, mais perçoivent également ces actions comme informatives. De plus, que les actions combinées peuvent être apprises avec un cadre RL en intégrant le raisonnement sur les observateurs potentiels pour enrichir les actions avec des aspects sociaux. Bien que la recherche présentée dans la thèse se limite à des cas spécifiques, elle démontre le potentiel prometteur des actions combinées dans les milieux HRI.

mots-clés: interaction homme-robot, apprentissage par renforcement, apprentissage robotique interactif, apprentissage à partir de la démonstration, communication sensorimotrice, lisibilité

Acknowledgement

Contents

List of Figures

List of Tables

List of Acronyms

AT Algorithmic Teaching.

DMP Dynamic Movement Primitives.

GUI Graphical User Interface.

HRI Human-Robot interaction.

INSEAD INSEAD-Sorbonne Université Behavioural Lab.

IRB Institutional Review Board.

IRL Inverse Reinforcement Learning.

LfD Learning-from-Demonstration.

MDP Markov Decision Process.

MT Machine Teaching.

ProMP Probabilistic Movement Primitives.

RL Reinforcement Learning.

SMC Sensorimotor Communication.

ToM Theory of Mind.

TP-GMM Task-Parameterized Gaussian mixture models.

VAE-DMP Variational Autoencoded Dynamic Movement Primitives.

Part I

Introduction

Chapter 1 Introduction

Contents

1.1 Motivations

The vision of creating artificial beings that can interact with humans and their environment like real humans is at least as old as human civilization. Nowadays, we refer to these artificial beings as robots, originating from the Czech word *robota* meaning 'forced labour'.

While the world pictured in modern science fiction (e.g. Asimov, 1988) is often heavily populated by different types of robots ranging from small little helpers to sophisticated humanoid robots, the idea of robots is actually much older. The idea already appears in classical Greek mythology where Hephaistos - the God of metallurgy and crafts - created automatons to work for him. At the end of the 17th century clockmaker build purely mechanically automatons with

a clockwerk. These automatons where not quite able to work yet, but were already able for example to write, draw or even perform magic tricks.

In modern times we progressed significantly from these automatons performing rather simple tasks to much more sophisticated robots. Nowadays robots are especially widely deployed in industrial environments to automate processes. Yet, we have not seen the wide deployment of robots "in the wild", namely the deployment of robots in unknown environments and structures. This deployment is difficult, since it is infeasible to program robots to cope with every possible scenario they might encounter beforehand.

From this fact arises the need to equip robots with the capability to learn how to solve new tasks. Another reason why robots have not been deployed in our day to day life is that they still lack the functionality to properly interact with humans in various situations.

In order to get to these capabilities it is useful to take inspiration from humans, as technology often has been inspired from nature. Robotics in particular has hugely been influenced by humans (appearance, behaviors, sensory input processing, etc.).

One crucial difference between humans and other species is the capability to share goals and intentions (Tomasello et al., 2005). This idea, namely the capacity of attributing a mental state to other people i.e. to infer observable beliefs, desires and intentions and interpret actions in relations to these mind state is called Theory of Mind (ToM) (Meltzoff, 1995; Dennett, 1987; Csibra and Gergely, 2007). ToM plays an important role in human communication.

Usually, humans communicate intentionally. In this context intentionally means that their goal is not to only say a certain thing, but rather that their communication partner understands a certain thing. This specific thing might be literal meaning of what is being said, but might as well require the communication partner to infer the meaning from what is being said and the global or specific context.

One special case of communication that is quite relevant to speed up a learning process is teaching. Humans adapted to teach ideas to conspecifics (Csibra and Gergely, 2006) and it's a powerful mechanisms for humans to learn

new skills that has not yet been extensively investigated in the context of Human-Robot interaction (HRI).

Another aspect important to communication is how the communication channels are used. The channel is the medium used to transmit the signal from source to destination (Shannon and Weaver, 1949). These channels could either be task channel or social channel (Sigaud et al., 2021). When for example humans teach to each other how to solve a certain task they can use these different channels to communicate what to do. When they show how to solve the task in question, this corresponds to using the task channel to communicate how to solve the task. Additionally they can explain how to solve the task using speech, here this explanation corresponds to using the social channel.

However, using different dedicated channels as task channel and social channel is not the only possibility to communicate something. It is also possible that only one channel is used to serve as combined task and social channel, meaning that one channel fulfills both their respective purposes. If humans, for example, teach a language they could use speech to teach the pronunciation of new vocabulary, but also to explain in which context it could be used.

If they teach how to solve a sensorimotor task this mean could be achieved by exaggerating certain aspects that are either particularly relevant to the task or were previously not executed by the learner. This leads us to the concept of Sensorimotor Communication (SMC). Pezzulo, Donnarumma, and Dindo (2013) introduce SMC as a communication that uses the same channel to execute an action and additionally convey information. The sensorimotor channel is of special interest for robotics, since the capability to interact with the real world is the most striking difference that differentiates robots from virtual agents. However, these actions that combine fulfilling a certain task and communicate additional information have not been extensively researched in the context of HRI.

1.2 Research Approach

Problem and Approach

Robots learning on their own by using different machine learning techniques like Reinforcement Learning (RL) is not new and has resulted in robots learning a variety of tasks, e.g. autonomous helicopter flight (Bagnell and Schneider, 2001), cart-pole swing-up (Deisenroth and Rasmussen, 2011) and jumping behavior for a robot dog (Kolter and Ng, 2009). However, the approach of humans explicitly teaching robots how to solve new tasks is less well developed.

The closest approach that has a well developed research body is Learning-from-Demonstration (LfD) (Argall et al., 2009; Calinon, 2019) that even combines well with RL (Kormushev, Calinon, and Caldwell, 2010; Mülling et al., 2013). However, in LfD the demonstrator is only demonstrating how to solve a task, solving the task them self. Usually, teaching something includes more than just solving the task in question. In this context we could also imagine that the teacher is including additional information. This aspect has been neglected so far, yet there is a difference between solving a task and teaching how to solve a task (Ho, Littman, MacGlashan, et al., 2016). This difference can be used to speed up the learning process (Ho, Littman, Cushman, et al., 2018). Therefore it's quite worthwhile to investigate this difference further. The difference between solving and teaching in HRI is the first aspect we address in our research.

The aspect that humans have intentions in their communication that goes beyond what is literally been communicated, has widely been neglected in research on robot learning. If we want to imitate human capability it is necessary to integrate the fact that both interaction partners usually have intentions. Interpretation of the intentions requires the interaction partner to mutually maintain a ToM. One impactful intention is the intention to teach. Further, in the context of robotics, the sensorimotor channel is of great importance since the practical values of robots stems from the fact that they can interact with the real world. Therefore, in this thesis, we focus on teaching intentions in combination with SMC. A simplified communication model we assume for our research is shown in Fig. 1.1.

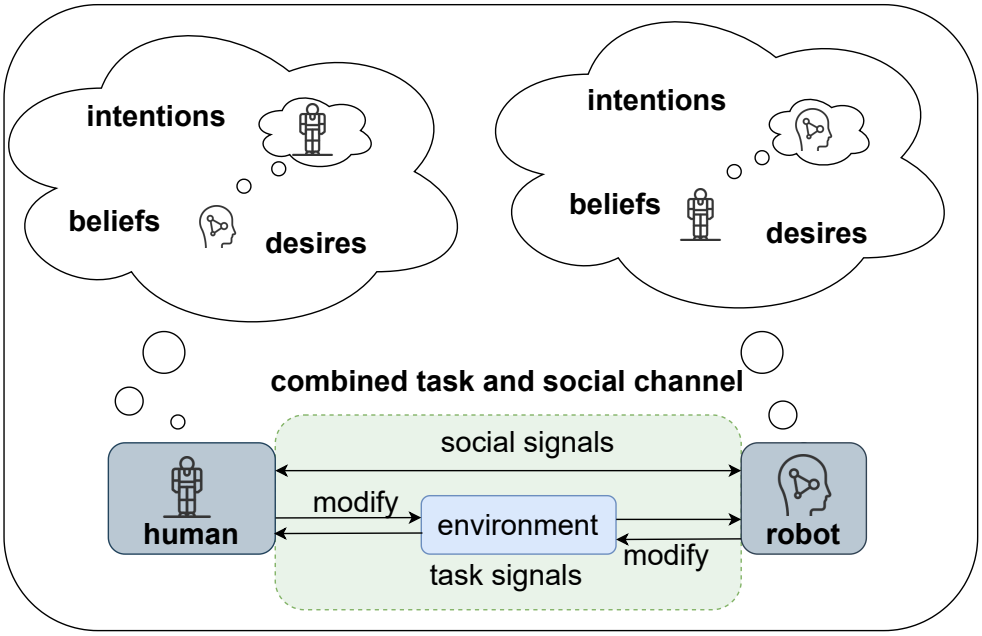


Figure 1.1: Simplified communication model integrating ToM and focusing on a combined task and social channel. We investigate how the task- and social channel characteristics of the combined task and social channel are used by humans when teaching robots. Furthermore, we investigate how these characteristics can beneficiary be used in HRI, and in particular in robot learning.

Research Questions

In our work we identified the following research questions that we address in this thesis:

1. Do humans make use of social channel characteristics when teaching robots a sensorimotor task? (Q1)
   (a) Does human behavior change when teaching a robot how to solve a task as opposed to just solving the task? (Q1a)
   (b) Do humans perceive this teaching behavior as more informative than the the solving behavior? (Q1b)
2. Are negative demonstrations useful to enrich approaches that use demonstrations to learn? (Q2)
   (a) Do humans perceive negative demonstrations as informative?
3. How can we integrate actions that make use of social channel characteristics into RL? (Q3)

1.3 Thesis Outline

This thesis consists of nine chapters. In the first chapter we introduce the domain of our research and our research approach. The next four chapters present the background and related research that relates and inspired the work in this thesis. The next chapter describes the general and specific communication model we assume for our research. The following two chapters describe the research we have implemented. The last chapter discusses our contribution and gives perspectives.

In the second chapter we present fundamental topics of cognition and communication that are, due to its multidisciplinary nature, important concepts for HRI in general and in particular for the work we present in this thesis. These include approaches to conceptualize communication processes between between humans and machines and insights on human reasoning in communication processes.

In the third chapter we give an overview of approaches to robot learning, and present important approaches like LfD and RL in more detail. Further, we discuss the particularities that distinguish robots from (virtual) agents.

In the fourth chapter we present approaches that explicitly focus on teaching. We present approaches that address formalizing teaching with mathematical frameworks, as well as research focusing on HRI settings where a human teaches a robot.

In the fifth chapter we introduce the idea of observer related metrics. The focus lies on predictability, and especially legibility, as these are two interesting, but fundamentally different metrics. Furthermore, we use legibility in our implementation in Chapter 8.

In the sixth chapter we propose a general communication model for HRI settings allowing for task signals, social signals, as well as combined task and social signals. Further, we present our specific model for communication focusing on combined signals, and how this model relates to our implemented research.

In the seventh chapter we present our user study investigating the difference of human behavior when they solve a sensorimotor task in contrast to teaching how to solve the task to a robot. In this chapter we address our first two research questions (Q1 & Q2).

In the eighth chapter we address the third research question (Q3). We present our RL base framework that we augmented with observer feedback in order to learn actions that communicate additional information to a potential observer at the same time.

In the last chapter, we conclude this thesis by discussing our research, the aspects that could be improved and give a perspective on future work.

1.4 Contributions

The three main contributions presented in this thesis can be summarized as follows:

- As a first contribution we present a model for communication that is, while still technical, better suited to account for human behavior in HRI than, for example, the code model (see Chapter 6). While we do not claim novelty of the ideas used in the model, we argue that having an explicit model is useful to position and guide further research focusing on communication aspects in HRI.
- The second contribution consists of insights on human behavior while teaching a robot how to solve a sensorimotor task, as well on human perception on this behavior (see Chapter 7). We show that when humans teach a robot how to solve a sensorimotor task, they communicate additional information via their actions as opposed when they just solve the task. We show further, that humans perceive these actions as more informative.
- Our third contribution is the proposition of a framework that integrates reasoning of a potential observer into the RL approach. We use the framework to compare different approaches implementing the reasoning process of the observer and show that they achieve higher legibility values than a classical RL baseline.

1.5 Publications

The results in this thesis have been published in the following publications:

- Chapter 7: Bied, Manuel and Mohamed Chetouani (2020). *"Exploring the Difference between Solving and Teaching in Sensorimotor Tasks"*. In: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. HRI '20. Cambridge, United Kingdom: Association for Computing Machinery, pp. 139–141. ISBN: 9781450370578. DOI: 10.1145/3371382.3378284.
- Chapter 8: Bied, Manuel and Mohamed Chetouani (2020). *"Integrating an Observer in Interactive Reinforcement Learning to Learn Legible Trajectories"*. In: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 760–767. DOI: 10.1109/RO-MAN47096.2020.9223338.

1.6 The Animatas Project

The research in this thesis was conducted within the EU project ANIMATAS. The aim of the project was to advance intuitive human-machine interaction with human-like social capabilities for education in schools. For this 15 early-stage researchers (ESRs) conducted research on three main research topics:

1. Perception
2. Social learning
3. Personalized adaptation

The research in this thesis can be localized within the social learning topic. While we do not propose applications that can directly be used in schools to foster learning, pedagogical situations (see Section 4.2) play an important role in our research. Thus, our contribution is more theoretical focused work that might be extended in the future to more practical applications that can be used in schools. For further information on the project please frequent the website at `http://www.animatas.eu`.

Funding Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955.

Part II

Background and Related Work

Chapter

2 Cognition and Communication

Contents

2.1 Introduction

HRI as a research field lies at the intersection of multiple disciplines like cognitive science, robotics, A.I. and social science (Kennedy et al., 2021). A variety of interaction protocols and algorithms in HRI have been influenced by concepts and ideas coming from fields that originally did not have any application with robotics in mind. While robotics has always drawn inspiration from nature (e.g. humanoids and bionics) often it is less important, if the robotic implementation behaves as the original, as long the implementation achieves what it is supposed to do. Often computer science and robotics can also inform social- and cognitive sciences (Brooks et al., 2002; Chaminade and Cheng, 2009; Wykowska, Chaminade, and Cheng, 2016; Sciutti et al., 2015).

In particular research on cognition and communication appears to be an important factor influencing HRI. In this chapter we present concepts of cognition and communication that are relevant to HRI in general and in particular to this thesis.

We start with the code model in Section 2.2, as it provides a great conceptualization of a communication process that is not only useful to understand communication better, but also to implement communication between electrical devices or a human and an electrical device as a robot.

Next, we continue with ToM in Section 2.3. ToM is an important foundation for more sophisticated communication and interaction in general. Furthermore, one important role within ToM plays goal attribution. This research motivates the metrics we present in Section 5, particularly a metric called legibility that we use in on of our experiments Chapter 8.

One essential part of the code model is the channel. In social interactions it is useful to differentiate between different channel types. In this context the social- and task channel seem to be particularly useful, as we present in Section 2.4. This differentiation plays also an important role in the general model we propose in Chapter 6.

Next, in Section 2.5 we present the concept of ostensive-inferential communication that introduces inference as an explicit factor into communication and addresses a major shortcoming of the code model.

SMC, as we present in Section 2.6 combines characteristics of task- and social channel in one channel proving a promising way to enrich communication in HRI.

2.2 The Code Model

The code model provides a conceptualized way to describe communication and had a great influence on technical implementation of communication systems. It was introduced by Shannon and Weaver (1949), and is also known as the Shannon-Weaver model.

A code enables two information-processing devices (organism or machine) to communicate. This mean is achieved by pairing messages with signals. A signal

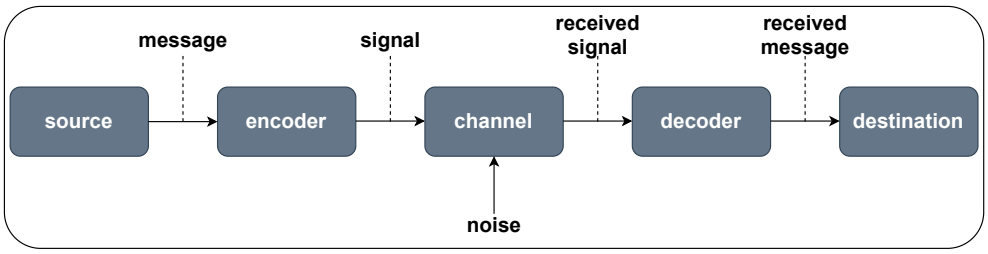


Figure 2.1: The code model after Shannon and Weaver. A message is transmitted from source to destination. The message itself can not travel, it has to encoded into a signal. The signal travels over the channel. The decoder reconstructs the original message from the signal. Recreation from (Sperber and Wilson, 1995).

is the modification of the environment by one device and recognizable by the other. A message is an internal representation to the device (Sperber and Wilson, 1995).

The model consists essentially of five parts:

1. An information *source* originates a message that is internal to the communicating devices. The message can not directly travel.
2. The *encoder* transforms the message into a signal that can be transmitted trough the channel.
3. The *channel* is the medium used to transmit the signal from source to destination. The channel can be influenced by noise disturbing the signal.
4. The *decoder* reverses the operation performed by the encoder to reconstruct the original message from the signal.
5. The *destination* is the information-processing device (organism or machine) for which the message is intended.

The code model provides a great technical description to implement a communication system, but can not account for the full range of human communication. An important factor missing is inference. While in rather artificial conditions inference can mimic encoding and decoding can mimic inference, these two types are essentially distinct, as Sperber and Wilson (1995) put forward. The concept of ostensive-inferential communication, described in Section 2.5, provides solutions to the shortcomings of the code model by introducing inference as an explicit factor.

While for electrical devices the encoding and decoding process can be explicitly implemented, it is clear that humans are no machines and these processes are not as explicit. Thus, in order to advance HRI it is necessary to better understand human behavior in interactions on the sending and the receiving end of communication. Research on ToM, as we see in the next section, can provide useful insights to better understand this behavior. In Section 6.3 we present a model that moves toward better integrating insights on human behavior.

2.3 Theory of Mind

ToM (also referred to as folk psychology) is the capacity of attributing a mental state to other people, reason about them and respond to their mental state (Meltzoff, 1995; Premack and Woodruff, 1978; Baker, Saxe, and Tenenbaum, 2009). This capacity includes inference of unobservable beliefs, desires and intentions and interpret actions in relations to these mind state. Fig. 2.2 illustrates mutual ToM of two people. Young children demonstrate these relatively sophisticated strategies by the age of five (Gergely et al., 1995). These sophisticated strategies develop over time, while less sophisticated subskills contributing to a ToM already develop earlier.

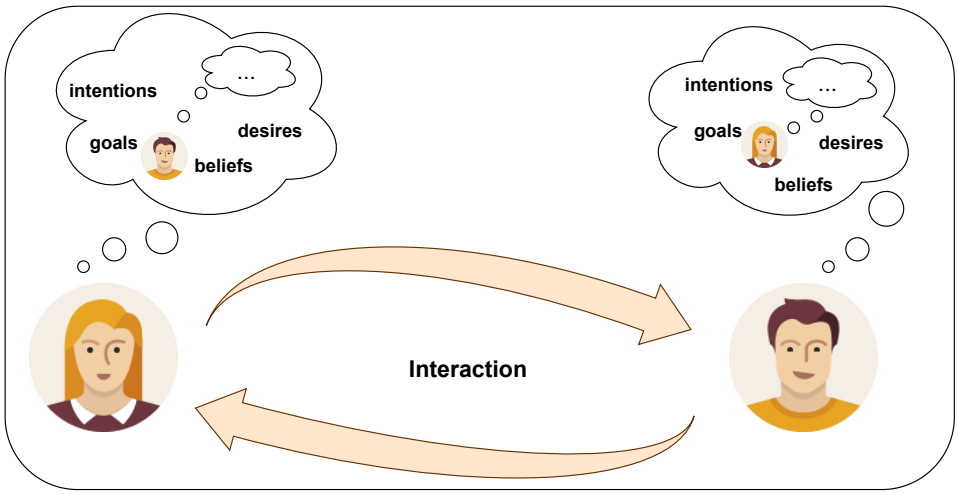


Figure 2.2: Two people having a mutual theory of mind. They mutually reason about the mental state of the other person and interpret actions in relation to these mind states.

One important aspect is the 'intentional stance' (Dennett, 1987). Intentions link desires and beliefs to actions. The object of an intention is always an

action, as opposed to desires that can have outcomes as objective (Malle and Knobe, 1997). The intentional stance approaches the explanation of other agents behavior by attributing intentional states (beliefs, desires, goals) as causes for their actions. The intentional stance can be found in quite young infants, i.e. Gergely et al. (1995) show that infants at the age of 12-month can already take the intentional stance. One important special case of intentional states can be found in teaching situations where the teacher intents the learner to understand a certain concept. We will have a closer look onto these situations in Section 4.2. However, as Csibra and Gergely (2006) put forward, while teaching (and learning) is assisted by ToM, the ability to teach is a primary ability that does not depend on ToM.

	Type of inference	
Primary function	Action-to-Goal	Goal-to-Action
On-line Prediction	Goal prediction: Predicting the likely effect of an ongoing action	Action anticipation: Predictive tracking of dynamic actions in real time
Social Learning	Discovering novel goals and artifact functions	Acquiring novel means actions by evaluating their causal efficacy in bringing about the goal

Table 2.1: Types of goal inferences and their respective functions. Recreation from Csibra and Gergely (2007).

Another important aspect within ToM are goals. Goals play a particular important role, since humans interpret observed behaviors as goal-directed actions (Meltzoff, 1995; Johnson, 2000; Csibra and Gergely, 2007; Carter, Hodgins, and Rakison, 2011). While goal-directed action understanding does not require any knowledge about the actor's mental state, these two types of knowledge are integral parts to form a building block for intention understanding (Carter, Hodgins, and Rakison, 2011). Csibra and Gergely (2007) identify two basic functions of goal attribution. Firstly, goal attribution allows for goal prediction and action anticipation. Secondly, goal attribution can enable long-term social learning.

Csibra and Gergely (2007) further identify two types of inference: "action-to-goal" and "goal-to-action". The first type of inference, the action-to-goal

inference, answers the question *What is the function of this action?* by predicting a future goal state from interpreting an ongoing action. Whereas the second type of inference, the goal-to-action inference, answers the question *What action would achieve that goal?*.

We see that intentions play an important role in human interaction, thus we integrate them in the model we present in Section 6.3. Further, in Chapter 8 we use the legibility metric (Section 5.2) measuring action-to-goal inference to evaluate how well an observer could interpret learned actions of an agent.

2.4 Social- and Task Channel

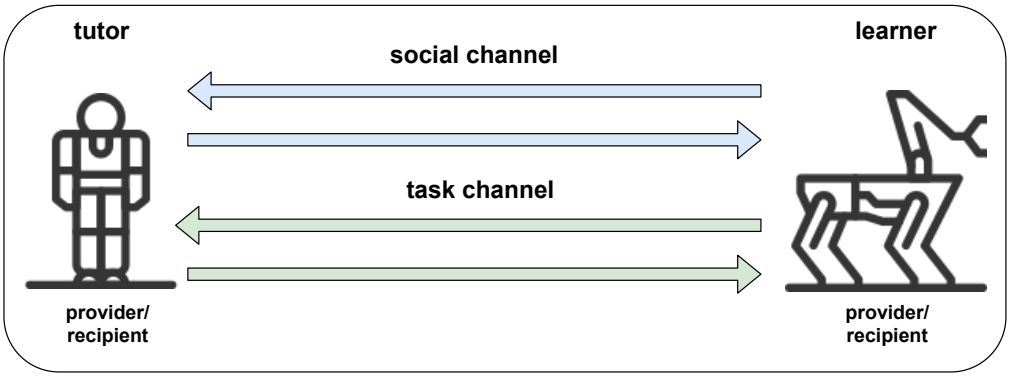


Figure 2.3: Mutual signal exchange using two channels. The task channel transmits task signals, e.g. demonstrations. Both participants can provide as well as receive both kind of signals. The social channel is used to transmit social signals such as feedback, requests, commitment signals, etc. Recreation from SIGAUD et al. (2021).

According to the Shannon-Weaver model (see Section 2.2), the channel is an essential part of communication. While the channel definition of the Shannon-Weaver model is rather technical, it is also useful in a social context. Thus, it is useful to have a closer look onto the properties of the available channels in a human robot interaction, in particular in a learning setting. In this section the focus is more on a theoretical point of view of the channels.

SIGAUD et al. (2021) conceptualize social learning processes (BANDURA and MCCLELLAND, 1977), more specifically interactions between a tutor and a learner as a mutual exchange process using two main communication channels. These two channels are the social channel and the task channel as shown in

Fig. 2.3. The exchange of signals is mutual, both participants can send and receive signals on both channels. In order to use both channels to its full extend, the tutor and the learner need to maintain a mental model of the interaction partner as described in ToM. (see Section 2.3).

Interaction from the task channel includes recognizing pedagogical signals (see Section 4.2), learning from demonstration (see Section 3.5), observational learning (Varni et al., 1979; Meltzoff, 1999; Burke et al., 2010) and inference from indirect goal-related signals (Bobu et al., 2020; Reddy et al., 2021). Except for tasks that explicitly include speech these signals are nonverbal.

Interaction from the social channel include feedback, instructions (see Section 3.3) and gaze (Nomikou et al., 2016; Fournier, Sigaud, and Chetouani, 2017).

Even if it is not necessarily called a channel, the idea of relying on task- as well as on social dimensions in an interaction is present in other work too. For example in Castellano, Pereira, et al. (2009) children play chess with a robot companion called iCat. The user engagement is detected using task features (e.g. the game state) as well as social features (e.g. the user smiling or looking at the iCat and the iCat displaying affective reactions). Similarly, in Leclère et al. (2016) task- and social features are used to distinguish, in the context of a mother-infant interaction, dyads that are at high-risk of showing neglect from dyads that are at low risk. In Ivaldi et al. (2014) related ideas are used in an interaction of a human with an iCub, a humanoid robot. In this work a social cue (gaze) is combined with task information (color of an object) to teach the color of objects to the robot. The participants indicated that they would like to see improved behaviors even if they were not task related.

A concept that explicitly combines pedagogical intentions on the task channel is SMC (see Section 2.6). Interaction from the social channel includes learning from feedback, instructions, joint attention and engagement. These signals can be verbal, i.e. feedback, or non-verbal like joint attention by gaze following. While research on robot learning usually uses these channels, research explicitly on the channel usage is rather sparse.

While not explicitly calling it a channel, Ho, Littman, Cushman, et al. (2015) and Ho, Cushman, et al. (2019) provide research on how people use evaluative

feedback, namely rewards and punishment. They investigate if evaluative feedback should be interpreted as reinforcement to shape the learner or as communication to signal to the learner to reason about the tutor's pedagogical goals. They come to the conclusion that people have a strong bias to use evaluative feedback as communication rather than as reinforcement.

Interesting insides on how people use the available channel for giving feedback to robots can be found in Thomaz and Breazeal (2008) and Thomaz and Breazeal (2006b). These works introduce the *Sophie's kitchen* framework, where people were asked to teach a reinforcement agent how to bake a cake. The works show that people are using the reward channel not only for rewards, but also for future directed guidance. The introduction of an explicit guidance channel speeded up the learning.

We see that social- and task dimensions play an important role in HRI research and we integrate this idea in our work in form of the task channel, social channel, and combined task and social channel (see Section 6).

2.5 Ostensive-Inferential Communication

As already explained in Section 2.2, the code model fails to account for the full range of human communication. The main defect of the code model is its descriptive inadequacy: there is more to communication than coding and decoding.

A good example for that is language. The same sentence can express a variety of thoughts, depending on the context and the relation between the communication partners. Comprehension of what is being said requires more than just the decoding of a linguistic signal. Take for example the utterance *"Do you know what time it is?"* could be a genuine question to get to know what time it is. The utterance could likewise be used to express that the speaker is quite annoyed about being called very late in the night. Thus, the meaning of the utterance depends on the context.

Grice (1957) provides an analysis that can be used as starting point for an inferential model of communication (Sperber and Wilson, 1995): *"[S] meant something by x' is (roughly) equivalent to '[S] intended the utterance of x to produce some effect in an audience by means of recognition of this intention".*

Thus, the communication succeeds when the hearer not only infers the linguistic meaning, but also what the speaker wants to convey.

While this approach has been criticized (e.g. Searle, 1969) that understanding intention can be just included in the decoding step, this critic misses an important point: communication can happen without any code (Sperber and Wilson, 1995). For example if Mary asks Bob how he is doing and he shows her his packet of painkillers in response. The conveyed message is that he is in pain, without the explicit presence of a code. Thus, the concept of inferential communication can be used for cases the code model can not account for. Consequently, the two models complement each other.

Sperber and Wilson (1995) extend the concept of inferential communication to ostensive-inferential communication. They put forward that ostension provides two layers of information: the informative intention and the communicative intention.

The informative intention is the information itself that has been pointed out. The communicative intention is to mutually manifest that the communicator has the informative intention. Thus, the communicative intention can be seen as "meta" intention. While in some cases the informative intention could be recognized without recognizing the communicative intention, in general failing to recognize the communicative intention might lead to missing relevant intention.

In the next section we will turn to SMC, a certain type of ostensive-inferential communication that we consider promising for HRI and we investigate further in Chapter 7.

2.6 Sensorimotor Communication

A specific type of ostensive-inferential communication that we identify as particularly interesting for HRI is SMC. While human communication has been the focus in many different disciplines. Several studies have focused on verbal and non-verbal communication like linguistic, gesturing and facial expressions. All these communication forms have in common that the channel used for communication is different to the channel used for execution of the action.

In contrast to these other forms of communication, in Pezzulo, Donnarumma, and Dindo (2013), the authors use the term SMC for a communication that uses the same channel to execute an action and additionally convey information. Thus, to explain it differently, SMC uses the sensorimotor channel as task- and social channel.

Later work (Pezzulo, Donnarumma, Dindo, et al., 2019) provides the working definition of SMC as "*signal that has a dual nature, and which combines a pragmatic action and a communicative action*". In Pezzulo, Donnarumma, and Dindo (2013), the authors formalize signaling in a computational framework in terms of parametrizable deviations for the optimal trajectory in order to be informative about the action choice while still achieving its pragmatic goal. They define signaling as the process of altering one's own behavior to facilitate its recognition by other persons.

For this they introduce probabilistic models, whereas the signaling distribution should be as close as possible to the original distribution, while at the same time offering a high discriminating power.

The starting point of the approach is inspired by a theory in computational motor control that each particular instantiating of an action can be associated to an internal model in the central nervous system (Wolpert and Ghahramani, 2000; Shadmehr, Smith, and Krakauer, 2010). In this approach, each model m_i is associated to a goal-directed action (e.g. reaching for an object to the left or right) mapping to a probabilistic trajectory. The evolution of a model is represented as $p(x_t|m_i)$ with x_t as the state of the system at time t, the entire sequence of states of the system resulting in following model m_i is denoted as $p(x|m_i)$. The perceiver's goal during interaction is to infer which model $m_{i_{ML}}$ has most likely generated the observed data where the index of the most likely model i_{ML} is given with:

$$i_{ML} = argmax_i p(m_i|x_{1:t}), \quad i \in 1, 2, ..., n \tag{2.1}$$

with n the number of available models. In this interaction, the performer's task is to facilitate this inferential process. Therefore, samples from the signaling distribution p^{sig} need a high probability of being sampled from the original distribution and a low probability of being sampled from a distribution belonging to another model. This mean can be achieved by using a modified rejection

sampling (Bishop, 2006) leading to the formal definition of the signaling distribution for the continuous case:

$$p^{sig}(x_t|m_i; w) \propto w_i \cdot p(x_t|m_i) \prod_{j \neq i} (1 - w_j p(x_t|m_j)/p_j^{max}), \tag{2.2}$$

with p_j^{max} the maximum value for the distribution $p(x_t|m_j)$ and w a weight vector modulating the contribution of individual models to the signaling distribution. Finding the signaling distribution can then be seen as optimization problem where the weight vector needs to minimize the following:

$$w_i(t) = argmin_{w(t)}[KL[p_i^{sig}(w(t)), p_i] + \lambda S(\theta - p_i^{simulated})] \tag{2.3}$$

with:

- the Kullback-Leibler divergence between the signaling distribution and the original one $KL(\cdot, \cdot)$,
- a parameter to control the amount of signaling λ,
- the perceiver's posterior probability of correctly recognizing the model m_i denoted as $p_i^{simulated}$. This posterior requires the assumption that the internal models are mutually known.
- an experimentally fixed threshold used by the receiver during model recognition θ,
- the logistic function S.

This formulation also permits to modulate the amount of signaling during the task, e.g. only for the first part of the action. They evaluate this model in three experiments, the first two experiments on synthetic data sets and the third experiment on real human data.

Summarizing the first experiment on synthetic data shows that signaling permits the perceiver to recognize a performed action faster while choosing between two possible actions, the second experiment on real human data shows that their proposed computational framework can successfully model the behavior shown by the real humans. In the third experiment they extend the model to incorporate three possible actions instead of only two.

Dockendorff, Sebanz, and Knoblich (2019) published a comment that

compares Pezzulo et al.'s newer work (Pezzulo, Donnarumma, Dindo, et al., 2019) with the previous work (Pezzulo, Donnarumma, and Dindo, 2013), and criticizes that the newer work leaves out the aspect about how co-actors distinguish action that are used for purely pragmatic goals from actions that combine pragmatic and communicative goals. Furthermore, in the comment they argue that *"the key distinguishing feature is that actions combining pragmatic and communicative goals will always involve deviations from efficient action performance"*. This earlier work and the comment point in a direction which features to use to recognize communicative goals.

We see that SMC provides a rich human communication method. Further, the possibility to combine task signals with social signals on the sensorimotor channel seems to be a promising direction for HRI. Thus, in Chapter 7 we study how humans use SMC to teach to a robot how to solve a sensorimotor task.

Chapter

3 Approaches to Robot Learning

Contents

3.1 Introduction

The need for approaches that enable robots to learn new tasks is motivated by the fact that robots can not be programmed in advance to solve all possible tasks they could encounter. The relevant algorithms and approaches will not always be clearly distinguishable from approaches where only a virtual agent is learning. Robot learning will integrate common approaches from machine learning that are not necessarily dependent on a physical agent. However, robots come with their own problems and challenges making it worthwhile looking into the problem of learning robots as its own domain (Kober, Bagnell, and Peters, 2013).

One possibility to approach robot learning is to have the robot learn fully autonomously. The most famous implementation for this approach is reinforcement learning, an approach inspired by trial-and-error learning first observed

by Thorndike (1898) in animals. Reinforcement learning is an approach well covered in literature as we present in Section 3.4.

Another possibility is to have the robot learn in interaction which a human. Chernova and Thomaz (2014) identify useful design elements that should be considered when designing robots that learn from interacting with humans:

- **Social interaction**: How can social aspects of the interaction can be leveraged? Which social cues can be leveraged to aid learning? Which social cues are most informative for task learning and which social cues are favored by the user.
- **Motivation for learning**: How initiates the interaction, will all learning be directed by the human or does the robot have an intrinsic motivation.
- **Transparency**: In order to guide the learning process in the best way possible, the teacher needs to maintain a model of the learner's knowledge. Therefore, it is an important question how the robot can make it's internal state transparent to the teacher. This transparency could be achieved by mimicking human communication or by the use of artificial interfaces that are not part of natural human communication.
- **Question asking**: Asking questions is an integral part of human learning. How can we implement question asking for robots? How do we provide the possibility for the human to answer the question in a way the robot can interpret? How can the gained information used to improve the underlying model? If multiple questions could be asked, how to decide which question to ask?
- **Scaffolding**: Scaffolding is the process of breaking the learning of a new skill into simpler sub-skills. This process often allows for greater efficiency, since the sub-skills can be reused. How can the scaffolding process leveraged in an interaction with the user?
- **Directing attention**: In the context of learning, similar to feature selection in machine learning, attention directing can be used to focus learning. It is an essential mechanism that contributes to the learning process.
- **Online vs. Batch learning**: This choice reflects an important aspect of

the interactive learning protocol and determines the flow of the interaction.

This idea has been implemented with a variety of approaches and we present an overview of commonly employed approaches in Section 3.3.

Some of these approaches are designed with particularly robots in mind, however some are designed for learning agents in general. Robots are a specific type of agents with certain properties. They are embodied agents with the capability to interact with the real world. Before turning to the learning approaches, we discuss these properties in more detail in the next section (Section 3.2).

3.2 Robots as Embodied Agents

In this section we have a closer look on characteristics that define a robot, because these characteristics are important when addressing the question how we can enable robots to learn. Robots can be considered as a special type of agents. An agent is an identity that is capable of making decisions. While theoretically these decisions could be random, in most cases these decisions will somehow be based on information gathered by or provided to the agent. Rational agents will try achieve the best outcome based on some objectives. While there might be useful applications of non-rational agents, we will consider agents as rational.

Following this definition robots are certain type of agent - robots are embodied agents. Technically a robot could be considered as a (software) agent reading and processing information coming from sensors and controlling certain hardware. However, we will consider all these parts together as integral parts making up one robot identity. In this sense, a robot is more than just an agent that comes with its own problems and advantages.

The most obvious and striking difference between a robot and a virtual agent is the fact that the robot can interact with the real world. The capability of interacting and manipulating objects in the environments has been used to learn different interesting tasks such as locomotion (Jun Nakanishi et al., 2004), the game "ball in a cup" (Kober, Mohler, and Peters, 2008) and table tennis (Muelling, Kober, and Peters, 2010)). The capability to interact with the real world makes robots well suited to automate tiring or even dangerous tasks that otherwise would need to be executed manually by

humans. Thus, these kinds of robots are widely deployed in industry.

However, robots can not only be used for automation, but also in a social context. Thus, research on social robots, robots that are able to communicate and engage in social interactions with humans has recently got more attention (Fong, Nourbakhsh, and Dautenhahn, 2003; Bütepage and Kragic, 2017; Dautenhahn, 2007). While most striking, the capability to physically interact is not the only advantage of using robots over virtual agents. Already the physical presence of a robot can yield its advantages. The work of Leyzberg et al. (2012) shows that the use of a robot increased learning gains for a human learner in comparison with a pure virtual system. Furthermore, a review on social robots for education (Belpaeme et al., 2018) identifies three advantages of robots over virtual systems. The first two already mentioned advantages are the capability to interact with the real world and increased learning gains for the human learner. The third advantage is that users show more social behavior beneficial for learning.

While using robots over virtual systems comes with advantages, it also comes with its own challenges and problems. These problems can be of two different types. The first type contains problems that directly concern the hardware. The second type contains restrictions on the software. These restrictions derive from the fact that hardware is used, but does not concern the hardware directly.

One considerable aspect concerning the hardware directly is the financial aspect: robotic systems are usually considerably more expensive than virtual systems. Not only the acquisition cost are higher, but also maintenance, since robots are exposed to wear and tear. They can break and malfunction for mechanical reasons, and unfortunately they often do in inappropriate moments. Even if they function properly, conducting robot experiments is time consuming. Somebody has to be around to ensure a smooth execution and verify that nothing goes wrong. Ensuring that multiple experiments in a row have the exact same conditions is difficult, even more so running multiple experiments in parallel. Furthermore, depending on the robot, malfunctions can be physically dangerous to humans interacting with or operating the robot.

On the software side we have the problem that typical assumptions that are often made in machine learning do not hold in robotics. Usually, it can neither be assumed that the true state is fully observable nor that the data is noise free.

Also the high-dimensional continuous state and action space is rather large (Kober, Bagnell, and Peters, 2013). While it is possible to simulate the robot, it is quite unrealistic that the robot will match this behavior in the real world, as a consequence, the algorithms that are being used need to be robust with respect to models not capturing all details of the real system correctly (Kober, Bagnell, and Peters, 2013).

3.3 Overview of Approaches to Robot Learning

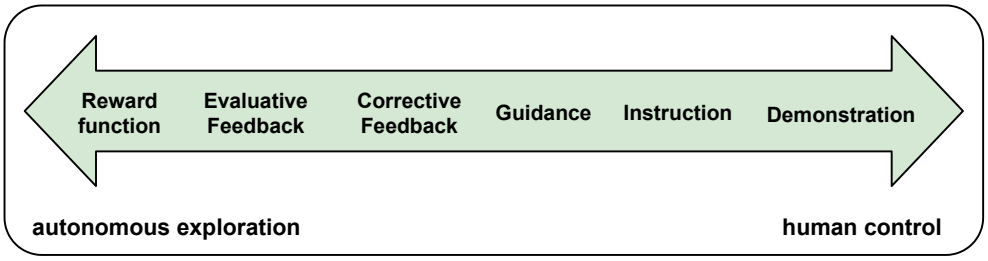


Figure 3.1: Approaches to robot learning can be located on a spectrum ranging from approaches where the robot learns fully autonomously to approaches where the human has full control of what is being learned. Recreation from Najar and Chetouani (2021).

After having a better understanding of the characteristics of a robot, we now turn to an overview of commonly used approaches to enable robots to learn. These approaches can be located on the exploration-control spectrum (Najar and Chetouani, 2021; Breazeal and Thomaz, 2008) as shown in Fig. 3.1. On the left side of the spectrum we find approaches where the agent learns autonomously like RL (Sutton and Barto, 1998). RL provides a mathematical framework to implement the idea of trial-and-error learning that has a broad corpus of research, particularly in robotics (Kober, Bagnell, and Peters, 2013). Classical reinforcement learning relies purely on the agent to explore the effects of its action on the environment. On this side of the spectrum the agent has a high autonomy and learns by itself.

When moving towards the right of the spectrum, the control influence of the human on the learning process increases. Coming from classical reinforcement learning we move to approaches that integrate feedback that the agent receives on taken action from a human tutor. These approaches are often combined with RL. However, how to integrate the feedback into the learning algorithms

needs research on its own (e.g. KNOX and STONE, 2012b; LI et al., 2019).

If we move further on the spectrum, we find guidance and instruction. These approaches limit the set of possible actions or suggest optimal actions (THOMAZ and BREAZEAL, 2006a). On the right corner of the spectrum we find the idea of demonstrations. This idea is implemented with the LfD framework (ARGALL et al., 2009; CALINON, 2019). The LfD approach is a commonly applied approach for robots learning new skills from humans, where the human demonstrator demonstrates how to solve a certain task to the robot. The robot learns from these demonstrations how to solve this particular task.

Except for classical reinforcement learning, all approaches on the spectrum can be counted toward interactive learning methods. In interactive learning approaches the teaching signals to an agent can be achieved via a variety of teaching channels like natural language (PALÉOLOGUE et al., 2018; CRUZ et al., 2015; KUHLMANN et al., 2004), computer vision (ATKESON and SCHAAL, 1997; NAJAR, SIGAUD, and CHETOUANI, 2019, computer code (MACLIN et al., 2005; TORREY et al., 2006, artificial interfaces (ABBEEL, COATES, and NG, 2010; SUAY and CHERNOVA, 2011; KNOX, STONE, and BREAZEAL, 2013) or physical interaction (AKGUN et al., 2012). NAJAR and CHETOUANI (2021) identify two main categories of teaching signals based on how they are produced: advice and demonstration. While these teaching signals could use the same channel, they are fundamentally different as the demonstration requires task execution and advice does not. In other words, demonstrations rely mainly (if not exclusively) on the task channel characteristics of the communication channel, while advice relies mainly on social channel characteristics (see Section 2.4).

Furthermore, NAJAR and CHETOUANI (2021) define advice as: "*teaching signals that can be communicated by the teacher to the learning system without executing the task*". Based on these considerations NAJAR and CHETOUANI, 2021 propose the following taxonomy of advice:

- **General advice** can be used to provide prior information on the task before the learning starts. It can be split into general constraints and general instructions.
- **General constraints** include information about the task such as domain concepts, behavioral constraints and performance heuristics.

- **General instructions** explicitly specify what actions to perform. It can either be provided in form of *if-then* rules or as detailed action plans.
- **Contextual advice** is provided during the task. It is dependent on the current state of the teacher-agent setting. It can be split into guidance and feedback.
- **Guidance** informs about future actions. In the most specific sense, it aims at limiting the set of all possible actions to a sub-set that is favored by the teacher.
- **Contextual instructions** are a particular type of guidance where only one action is suggested by the teacher.
- **Feedback** informs about past actions taken by the agent. It can be split into corrective and evaluative feedback.
- **Corrective feedback** can consist of either a corrective instruction or a corrective demonstration.
- **Evaluative feedback** can be provided in different forms. These include scalar values, binary values, positive reinforcer or categorical information. Also preferences between alternatives can be counted towards evaluative feedback.

It is difficult to clearly structure and separate all approaches used in interactive learning. For example, one problem with this taxonomy is that that Najar and Chetouani (2021) include demonstrations (e.g. Subramanian, Isbell, and Thomaz, 2016) in guidance. This makes sense, since demonstrations can be used to guide the learning process, however it does not go along with the previous definition of not executing the task. While this taxonomy is not perfect, it gives a good overview about applied techniques in the middle of the spectrum.

In the next two sections we present the two approaches on both ends of the spectrum, LfD (Section 3.5) and RL (Section 3.4).

3.4 Reinforcement Learning

RL implements an approach that is inspired by trial-and-error learning first observed by Thorndike (1898) in animals. An agent (the learner and decision maker) explores the space of possible strategies by interacting with the environment that comprises everything outside the agent's control. The agent receives feedback on the outcome of the chosen action. This information is used to improve the strategy and finally find the optimal strategy.

Markov Decision Process

The standard way of formalizing reinforcement learning problems is the use of a Markov Decision Process (MDP). The Markov property states that the current state information includes all relevant information concerning the environment. The next state after taking an action does only depend on the action and the current state and not, for example, on the history of previous states. While most of theoretical guarantees only hold if the requirements of the Markov property are fulfilled, many approaches work well in practice even if the problems do not fulfill these requirements (Kober, Bagnell, and Peters, 2013). Further, a MDP is defined as tuple $(S, A, \mathcal{T}, R, \gamma)$:

- S is the set of possible states (also called the state-space),
- A is the set of possible actions (also called the action-space),
- $\mathcal{T} : S \times A \times S \to P(s' \mid s, a)$ defines the state-transition probability function, with $P(s'|s, a)$ representing the probability that the agent transitions to state s' when taking the action a,
- $R : S \times A \times S \to \mathbb{R}$ defines the reward $r(s, a, s')$ that the agent receives when transitioning from state s to the new state s' while taking action a,
- $\gamma \to [0, 1]$ is the discount factor describing how much rewards for the recent decision are taking into account.

Interaction loop

The interaction loop for a RL problem (Fig. 3.2) can be described as follows. At each time step, the agent receives the current state $s_t \in S$. Based on the

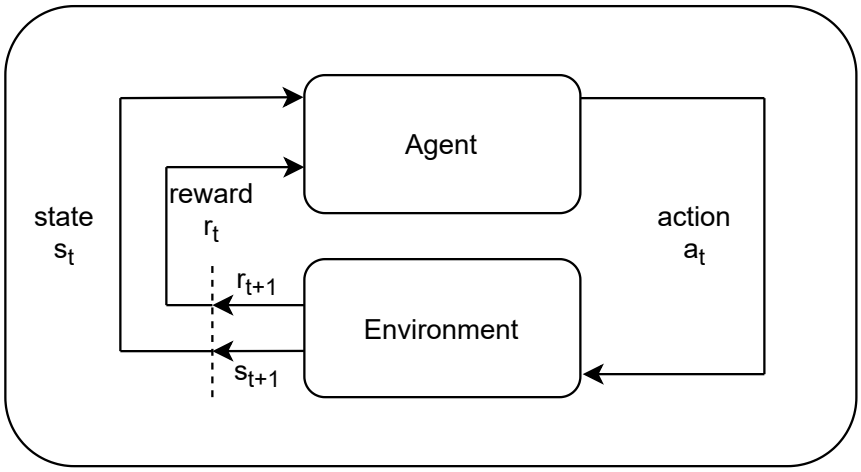


Figure 3.2: The interaction of the agent with the environment. The agent knows that the environment is in state s_t, takes an action a_t at the time step t and receives the corresponding reward r_t. The environment transitions to the next state s_{t+1}. Recreation from SUTTON and BARTO (1998).

current state it executes action $a_t \in A$. The environment transitions to the next state s_{t+1} and provides the reward r_{t+1} to the agent.

Agent's objective

The agent's objective is to maximize the cumulative received reward J while executing a certain policy π. A policy is a mapping from state to action that can either be deterministic or stochastic:

- $\pi : S \to A$ for deterministic policies and
- $\pi : S \times A \to [0, 1]$ for stochastic policies.

The cumulative reward for a certain policy is denoted as:

$$J = \sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, \pi(s_{t+k}), s_{t+k+1}). \tag{3.1}$$

Value functions

In order to evaluate a certain policy usually two value functions are used. These functions are called *state-value function* and *action-value function* and are used to evaluate how good a certain state is, respectively how good it is to perform certain actions in a certain state.

State-value function

The state-value function $V^{\pi}(s)$ is the expected return when the agent starts in state s and follows the policy π. There is a relationship between the value of state and the values of its successor states. This relationship can be expressed with the *Bellman equation* for $V^{\pi}(s)$ (Sutton and Barto, 1998):

$$V^{\pi}(s) = \sum_{a} \pi(s,a) \sum_{s'} P(s'|s,a)[r(s,a,s') + \gamma V^{\pi}(s')] \tag{3.2}$$

The agent now needs to find the optimal policy π^*, respectively one of the optimal policies if there are multiple optimal policies. The return of π^* is greater (or at least as great for multiple optimal policies) than all other policies. The state-value function can now be used to express the relationship between two policies, a policy π is better than or equal to a policy π' if and only if $V^{\pi}(s) \geq V^{\pi'}(s)$ for all $s \in S$ The optimal state-value function is defined as:

$$V^{*}(s) = \max_{\pi} V^{\pi}(s), \tag{3.3}$$

for all $s \in S$ leading to the *Bellman optimality equation*:

$$V^{*}(s) = \max_{a} \sum_{s'} P(s'|s,a)[r(s,a,s') + \gamma V^{*}(s')] \tag{3.4}$$

Action-value function

The value of taking an action a while being in state s and following the policy π can be described with the action-value function for policy π denoted as Q^{π}:

$$Q^{\pi}(s,a) = \sum_{s'} P(s'|s,a)[r(s,a,s') + \gamma V^{\pi}(s')] \tag{3.5}$$

Approaches to find the optimal policy

Approaches to find the optimal policy include value-based approaches, policy search and actor-critic methods. Value-based algorithms obtain the optimal policy by iteratively optimizing the value function. Popular approaches for value-based RL that are often used as benchmarks include Q-Learning (Watkins and Dayan, 1992) and SARSA (Sutton, 1996). The main difference between these two approaches is that Q-learning learns off-policy and SARSA learns

on-policy. Off-policy learning means that the policy being updated differs from the policy being followed. On-policy means that the algorithm estimates the value of the policy that is actually being followed.

A simple one-step Q-learning, where Q directly approximates the optimal action-value function Q^*, is defined by:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{3.6}$$

The step size α with $0 < \alpha < 1$ defines how strongly to move towards the new estimate at each iteration, the larger α, the larger the step towards the new estimate. The discount factor γ with $0 \leq \gamma \leq 1$ determines how strongly to take future rewards into account. When γ is 0 the agent will only consider current rewards and with increasing γ the agent takes future rewards more strongly into account. The SARSA algorithm only differs in the update function:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{3.7}$$

Different approaches can be used for deriving the policy π from the Q-function at decision time. Typical approaches include the ϵ-greedy and softmax action selection strategy. The ϵ-greedy strategy uses the exploration rate ϵ and selects the optimal action w.r.t. the Q function most of the time and with a small probability ϵ a random action:

$$a_t = \begin{cases} \max_{a \in A} \quad Q(s_t, a) \text{ with probability } 1 - \epsilon \\ \text{random action with probability } \epsilon \end{cases} \tag{3.8}$$

The softmax strategy assigns the highest probability to the optimal action w.r.t. the Q function and lower probabilities to all other actions. The lower the value of the action the lower the probability that is assigned to that action. A typical choice is a Gibbs (also called Boltzmann) distribution:

$$\pi(s, a) = Pr(a_t = a | s_t = t) = \frac{e^{Q_t(a)/\tau}}{\sum_{a' \in A} e^{Q(s,a')/\tau}} \tag{3.9}$$

where τ can be used to modulate how sharp the probability distribution peaks around the optimal value. Policy search approaches directly search in the policy space to find the optimal policy. Policy search deals better with

typical challenges encountered at robot reinforcement learning (e.g. high-dimensional continuous action space) (Deisenroth, Neumann, and Peters, 2013; Kober, Bagnell, and Peters, 2013). Policy search approaches can be divided into two main categories: evolutionary methods (Moriarty, Schultz, and Grefenstette, 1999) and policy gradient methods (Sutton, Mcallester, et al., 2000; Ng and Jordan, 2000).

A hybrid method of policy gradient and value-based methods offer actor-critic based approaches (Barto, Sutton, and Anderson, 1983; Grondman et al., 2012). These approaches combine advantages of both methods by learning a parameterized policy called the actor and the value function called the critic at the same time.

Shaping methods integrating advice

Advice can be integrated into RL systems at different levels. These levels are the reward function, the value function, the policy or the decision making. Thus, the four main strategies can be identified as reward shaping, value shaping, policy shaping and decision biasing (Najar and Chetouani, 2021).

An important model to mention in the context of interactive RL is TAMER (Knox and Stone, 2008; Knox and Stone, 2009). TAMER directly models the human rewards and myopically learns from this model. TAMER itself is not a RL technique. However, further work introduces an approach called TAMER+RL (Knox and Stone, 2010; Knox and Stone, 2011; Knox and Stone, 2012b) that integrates TAMER with RL. Furthermore, other models that integrate human feedback into RL use concepts borrowed from TAMER (e.g. COACH (Celemin and Ruiz-del-Solar, 2015; Celemin, Ruiz-del-Solar, and Kober, 2019)) or extend it (e.g. ACTAMER Vien, Ertel, and Chung, 2013).

Knox and Stone (Knox and Stone, 2010; Knox and Stone, 2011; Knox and Stone, 2012b) propose different approaches to integrate a human given reward $\hat{H}$ with traditional RL.

For reward shaping they propose to replace the reward with the sum of itself and a weighted human given reward defined as follows:

$$R'(s,a) = R(s,a) + \beta\hat{H}(s,a) \tag{3.10}$$

Note that this approach does not fulfill the requirement to be a potential-based reward function (Ng, Harada, and Russell, 1999), and can lead to positive circuits (Ho, Littman, Cushman, et al., 2015; Knox and Stone, 2012a).

Value shaping methods modify directly the Q-function instead of the reward. Q-Augmentation also proposed by (Knox and Stone, 2010; Knox and Stone, 2011; Knox and Stone, 2012b) modifies the Q-function as follows:

$$Q'(s,a) = Q(s,a) + \beta \hat{H}(s,a) \tag{3.11}$$

However, augmenting the Q-function like this can lead to convergence problems when used with evaluative feedback. This problem can occur since the Q-function also informs about the proximity to goal, while this information might not be included in the evaluative feedback (Najar and Chetouani, 2021; Ho, Littman, Cushman, et al., 2015).

Policy shaping (Griffith et al., 2013) does not manipulate the reward, but affects the policy directly. Action biasing and control sharing as proposed by (Knox and Stone, 2010; Knox and Stone, 2011; Knox and Stone, 2012b), fall into this category.

Action biasing also uses Eq. 3.11, but only during action selection and does not change the Q-function directly. Thus, the action is determined with:

$$a^* = \text{argmax}_x[Q(s,a) + \beta \hat{H}(s,a)] \tag{3.12}$$

Control sharing directly guides exploration toward human favored state-action pairs. The decision is taken according to $\hat{H}$ with the following probability:

$$Pr(a = \text{argmax}_a[\hat{H}(s,a)] \tag{3.13}$$

and according to agents action selection mechanism otherwise. β is used as threshold to set the probability.

Decision biasing is similar to policy shaping, however the policy is not corrupted and the advice is not modeled (Najar and Chetouani, 2021). Rosenstein and Barto (2003) use the provided instruction to bias the decision as fol-

lows:

$$a \leftarrow ka^E + (1-k)a^S, \tag{3.14}$$

with the actor's exploratory action, the supervisor's action and an interpolation parameter k.

We have now covered the approaches on the left to the middle of the exploration-control spectrum (see Section 3.3), we see that there is a variety of possibilities to integrate advice into RL that come with different advantages and disadvantages. In Chapter 8 we use reward shaping to integrate observer feedback into our RL framework and Q-Learning as baseline comparison.

In the next section we turn the approach that is located most to the right on the spectrum: LfD.

3.5 Learning from Demonstration

LfD, also referred to as Programming by Demonstration or Learning by Imitation, is a popular approach to transfer new skills to agents, and particularly robots, in a user-friendly and intuitive manner (Argall et al., 2009; Billard et al., 2008). LfD draws its inspiration from imitation learning in humans and animals. In this context, there are four important questions to answer: *what-to-imitate*, *how-to-imitate*, *when-to-imitate* and *who-to-imitate* (Nehaniv and Dautenhahn, 1999; Calinon, Guenter, and Billard, 2007).

LfD research usually focuses on the questions what and how to imitate. The question what to imitate corresponds to learning a skill and the question how to imitate corresponds to the encoding of the skill (Billard et al., 2008).

The typical LfD process can be summarized as follows (Pais Ureche and Billard, 2015): First, data is demonstrated and recorded. Subsequently, the data is analyzed and encoded into a model of the task. The last step is to execute the task while using the learned model of the task.

The first step, namely capturing the data can be achieved in various forms, for example with motion sensors (Steffen et al., 2010) or a visual motion capture system (Lioutikov et al., 2015). Demonstrations have a fundamental problem when using these forms, since the body of the teacher and the robot differ. This is known as the correspondence problem (Dautenhahn and

Nehaniv, 2002). In order to solve this problem a mapping from the body of the human to the body of the robot has to be found. Another approach to circumvent this problem is to directly execute the demonstrations on the robot. A popular approach for this is kinesthetic teaching (Calinon and Billard, 2007). Within kinesthetic teaching the teacher guides the motions of the robot directly to solve the task at hand. However, this process can get difficult when, for example, multiple joints of the robot have to be controlled at the same time.

The details of the second step, analyzing the data and encoding into a model, depends highly on the chosen model. Popular approaches are biological inspired basic elementary movements called movement primitives (Bizzi et al., 2002; Flash and Hochner, 2005). Popular implementations include Dynamic Movement Primitives (DMP) (Ijspeert, Nakanishi, and Schaal, 2002; Ijspeert, Jun Nakanishi, et al., 2012), Probabilistic Movement Primitives (ProMP) (Paraschos et al., 2013; Paraschos et al., 2018) and Task-Parameterized Gaussian mixture models (TP-GMM) (Calinon, 2015a; Calinon, 2015b). There are also approaches that combine movement primitives with neural networks like Variational Autoencoded Dynamic Movement Primitives (VAE-DMP) (Chen, Bayer, et al., 2015; Chen, Karl, and van der Smagt, 2016; Chaveroche et al., 2018).

The last step, the reproduction, corresponds to the *what-to-imitate* problem. Reproduction, requires a controller able to execute the learned model on the robot. The model needs to be robust to errors (i.e. tracking error) during execution. This step corresponds to the *what-to-imitate* problem.

Usually, the LfD approach either assumes an expert user that is giving (nearly) perfect demonstrations, or a novice (also called naïve) user that might give flawed demonstrations. While the work of Breazeal, Berlin, et al. (2006) presents an approach in the spirit of the saying "Do as I say, not as I do", this approach assumes a naïve user giving flawed or ambiguous demonstrations and provides a mechanism for the robot to infer the humans goals even if the human is not achieving these goals. However, the existing research usually assumes that the user is just solving the task, giving the best solution they are capable of.

There exists little research with the assumption that the demonstrator is not

only solving (doing) the task, but explicitly teaching the robot how to solve the task. Thus, in Chapter 7 we implement the demonstration phase of the LfD pipeline focusing on human teaching behavior. Since in our work we combine LfD with the idea of teaching, we present approaches that explicitly focus on teaching in the next chapter.

Chapter 4 Teaching Machines and Robots

Contents

4.1 Introduction

In this chapter we present approaches that explicitly focus on teaching that either provide a possibility to express the teaching process in a mathematical framework, investigate human teaching behavior towards robots (respectively virtual agents) or a mixture of both.

Note that approaches where a robot learns from human signals might not always be clearly separable from approaches where a human explicitly teaches a robot. In principal both approaches can be combined. However, since *"humans are adapted to transfer knowledge to, and receive knowledge from, conspecifics through teaching"* as Csibra and Gergely (2006) put forward in their *pedagogy hypothesis*, not benefiting from explicit teaching misses out on a great opportunity to improve approaches where robots learn from humans. Pedagogy plays an important aspect in human teaching as we present in Section 4.2.

The idea of teaching a robot, or a machine in general, is not new. Already Alan Turing expressed this idea as: *"It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. That process could follow the normal teaching of a child. Things would be pointed out and named, etc"* (TURING, 1950). One way to formalize the teaching idea in a mathematical framework offers machine teaching as we present in Section 4.3. This framework is also useful to evaluate human teaching behavior towards robot.

While research focusing on how humans are teaching robots is sparse (VOLLMER and SCHILLINGMANN, 2018), within the last few years there has been an increased research interest in this direction. We present this kind of research in Section 4.4.

4.2 Pedagogy

Pedagogy plays an important role in human social learning. As already pointed out in the section on ToM (see Section 2.3), intentions play an important role in human communication in general. A specific type of intentions important for social learning are pedagogical intentions. Pedagogical intentions are intentions that link desires and beliefs to actions that help the learner to acquire new knowledge.

Another important concept is *natural pedagogy*, a term introduced by CSIBRA and GERGELY, 2009 for a social communicative learning mechanism, where the knowledgeable teacher selectively manifests 'for' the learner the relevant information in order to acquire new knowledge. Natural pedagogy is a form of ostensive communication. For example, a demonstration becomes more than just the solution of a task, but can yield additional information about the task.

Similarly, SHAFTO, GOODMAN, and GRIFFITHS (2014) introduce the concept of pedagogical situations. These are defined as *"situations are settings in which one agent is choosing information to transmit to another agent for the purpose of teaching a concept."* In a pedagogical situation the (optimal) teacher selects the data to present to the learner maximizing its belief in the correct hypothesis. The authors formalize pedagogical reasoning as a Bayesian model.

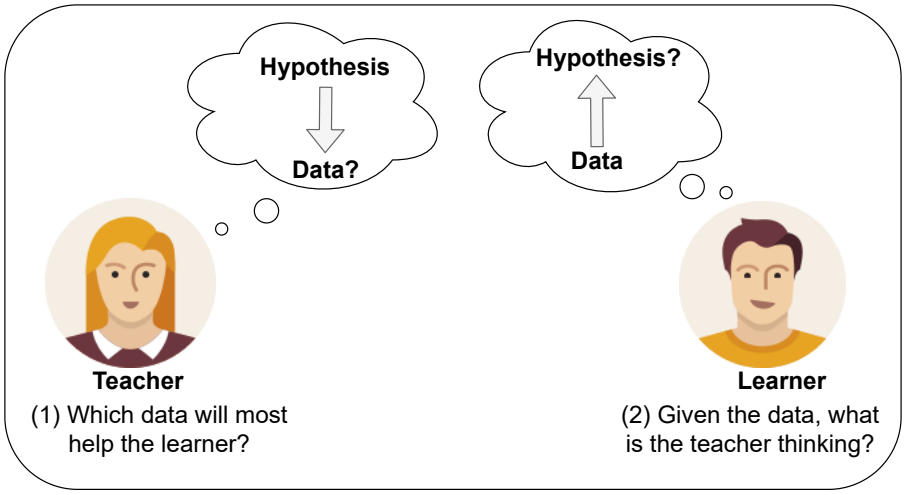


Figure 4.1: Schematic depiction of pedagogical reasoning. The teacher (left) knows the correct hypothesis and tries to choose data that is most helpful to the learner to infer the true hypothesis. The learner (right) observers the data and tries to infer from the data the hypothesis the teacher wants to convey. Recreation from SHAFTO, GOODMAN, and GRIFFITHS (2014).

While these equations requires recursion, pedagogical reasoning is the outcome of a psychological process that might not require explicit recursion. In this model the probability that the teacher selects the data d given the hypothesis h is given with:

$$P_{teacher}(d|h) \propto P_{learner}(h|d)^{\alpha} \tag{4.1}$$

and the probability that the learner infers the correct hypothesis h given the data d is given with

$$P_{learner}(h|d) = \frac{P_{teacher}(d|h)P(h)}{\sum_{h'} P_{teacher}(d|h')P(h')}. \tag{4.2}$$

The parameter α modulates how pedagogical the teacher chooses their examples, $\alpha = 1$ corresponds to a maximum of pedagogical intentions and $\alpha = 0$ to random sampling. These two equations are mutually dependent and plugging Eq. 4.2 into Eq. 4.1 leads to:

$$P_{teacher}(d|h) \propto \left(\frac{P_{teacher}(d|h)P(h)}{\sum_{h'} P_{teacher}(d|h')P(h')}\right)^{\alpha} \tag{4.3}$$

The model was tested with human participants in different experiments with a teacher and a learner. One of the experiments was on rule-based concepts.

In this experiment the teacher had to chose examples given a rectangle, and the learner had to guess the rectangle given the examples. There were three conditions: *Teaching-Pedagogical Learning, Pedagogical Learning* and *Non-Pedagogical Learning*. In the *Teaching-Pedagogical* condition the participants first had the role as teacher and as learner afterwards. In the *Pedagogical Learning* condition the participants had the role of the learner, but knew there was a (not present) teacher choosing the examples. In the *Non-Pedagogical Learning* only had the role as learner and knew that the examples where not selected by a teacher. The results show that their model predicts human data quite well. As predicted by the pedagogical model, people chose overwhelmingly examples in the corner as predicted for the *Teaching-Pedagogical Learning* and *Pedagogical Learning* condition. For the *Non-Pedagogical Learning* condition the examples were nearly evenly split between corners and non-corners. Nevertheless, the authors point out that the learning cases in their experiments are much simpler than cases that may be encountered in educational contexts. They see their work as first step to understand how learning is affected by pedagogical situations, and present a new framework to explore implications for education.

Similarly, Ho, Littman, MacGlashan, et al. (2016) investigate the difference between doing and showing. Showing corresponds to pedagogical situation and doing to a non-pedagogical situation. In their experiments, they used two conditions: in the *Do* condition they promised the participants a bonus based on their performance on the task, in the *Show* condition they promised the participants a bonus based on how well a randomly matched partner who was shown their response would perform on the task.

In the first experiment the task consisted in showing a trajectory from a start position to different possible goal positions in a 2-dimensional grid world. The participants in the *Show* condition tended to choose paths that disambiguate their goal as compared to the participants in the *Do* condition.

The grid world in the second experiment consisted of different colored fields. Each color indicated a different reward for passing this field. Some colors indicated dangerous fields that yielded a negative reward. The participants in the Do condition took the most efficient routes, the Show participants took paths that led through multiple safe field types, showing that their exists a difference between teaching and solving a task.

The authors further modelled the behavior in the two conditions by using similar Bayesian models as before presented. They show that an Inverse Reinforcement Learning (IRL) algorithm can beneficially learn from the *Show* condition. In Ho, Littman, Cushman, et al. (2018) they extend their previous work by integrating communicative goals of a showing demonstrator into the reasoning of the observer, improving the confidence of the model.

Milli and Dragan (2019) use the data of Ho, Littman, Cushman, et al. (2018) to investigate if it's safer to assume that the human behaves literal or pedagogical. The authors find that on the empirical data the assumption of a literal human achieves better performance even when people try to be pedagogic. They improve the performance in comparison to models that assume either a literal- or a pedagogical human model by introducing a mixture model integrating the literal- and pedagogical human model. However, they also state that the problem in assuming a pedagogical human are the unforeseen deviations of human behavior in the behavior model that render the model unstable and thus conclude it's safer to use the literal human model even when people try to be pedagogic.

We see that in the recent years there has been upcoming research to approach pedagogy from a computational point of view. Nevertheless, the applications are still quite limited and further research needs to be done in this area. Thus, in Chapter 7 we conduct a similar experiment as Ho, Littman, MacGlashan, et al. (2016) to investigate the difference between a solving a sensorimotor task and teaching the task to a robot.

4.3 Machine Teaching

Machine Teaching (MT) is an interesting field to look at in the context of Human-Robot Interaction and robot learning, since it offers a formal framework to capture a problem where a task or concept is taught to one agent by another agent. While the aim of machine learning is to optimize a model given a data set, the aim of machine teaching is to optimize a data set given the algorithm and the model that should be learned while minimizing the cost. This aim is very closely related to the aim of Algorithmic Teaching (AT) (Goldman and Kearns, 1995; Bengio et al., 2009; Cakmak and Lopes, 2012a), which is to minimize the number of demonstrations required to train an agent. However,

MT does not simply try to reduce the number of required demonstrations, but associates 'costs' to the teaching.

X. Zhu et al. (2018) give the following formal definition of MT:

$$\begin{aligned} \min_{D,\hat{\theta}} \quad & \text{TeachingRisk}(\hat{\theta}) + \eta \text{TeachingCost}(D) \\ \text{s.t.} \quad & \hat{\theta} = \text{MachineLearning}(D) \end{aligned}$$

Here D denotes the data set, $\hat{\theta}$ denotes the parameter that is learned by a given machine learning algorithm. The $\text{TeachingRisk}(\hat{\theta})$ describes the cost of the error between the learned model and the optimal model that should be learned, while $\eta\text{TeachingCost}(D)$ describes the weighted costs associated with providing the data set D to learn from. While these terms are generic in the definition, they offer a possibility to adapt them to the learning task.

AT usually follows two prominent threads: the teaching dimension (Goldman and Kearns, 1995) and curriculum learning. The teaching dimension is the smallest teaching set size to acquire the concept to be learned. Curriculum learning follows a strategy that starts with clear examples and continues with more ambiguous examples (Bengio et al., 2009).

Applied to a 1-dimensional discrete classification task with positive examples on one side of a decision boundary and negative examples on the other side, as it can be seen in Fig. 4.2, the teaching dimension approach predicts a set of two demonstrations, one left to the decision boundary and one right to the decision boundary. The curriculum approach predicts a strategy that starts with examples most to the right and most to the left and then gradually approaches the decision boundary.

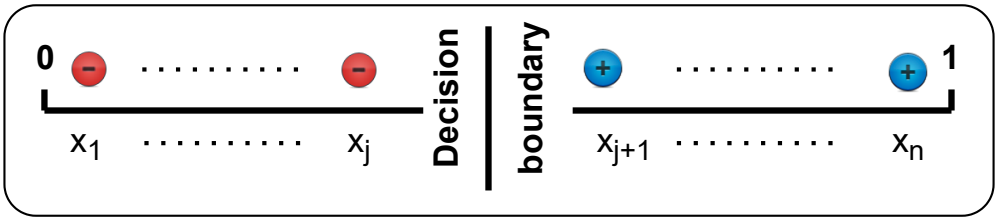


Figure 4.2: Exemplary 1-dimensional discrete classification task with positive examples on the right side and negative examples on the left side. Recreation from Khan, Mutlu, and X. Zhu, 2011.

In this thesis, we pickup on the idea of having costs of teaching in Chapter 7 by

introducing different conditions that only vary in the number of demonstrations the participants can give to teach a sensorimotor task to a robot.

4.4 Humans Teaching Robots

Vollmer and Schillingmann (2018) provide a review over studies presenting teaching interactions with a robot learner and a human teacher that also report on the human teaching behavior. This research is sparse and while the authors did not claim exhaustiveness, they only found 18 papers matching these criteria. While all of these papers studied teaching interactions, only in five (28%) of the studies the robot actually learned something. They mention two possible reasons: the high implementation effort of a suitable learning algorithm and the introduction of undesired variability into the study.

In Khan, Mutlu, and X. Zhu (2011), the authors investigate how humans choose examples to teach a task that corresponds to a 1-dimensional classification task (as explained in Section 4.3) to a robot. The task consists of ordering pictures of objects along a line of how graspable they are, and then provide examples to teach this graspability to a robot. The authors found the three following strategies: The extreme strategy that corresponds to curriculum learning (examples on both extreme sides), the positive only strategy, where people only gave positive examples and the linear strategy, where people moved from left to right (or vice-versa). Furthermore, the work mentions the boundary strategy, examples on both sides close to the decision boundary corresponding to the strategy predicted by the teaching dimension, however they authors could not find empirical evidence for this strategy.

Cakmak and Lopes (2012b) extends AT to an optimally teaching sequential decision tasks. In this work an IRL agent learns from human demonstrations. The authors find that the natural teaching behavior is normally sub-optimal, but that spontaneous optimal teaching is possible. Furthermore, they find that providing instructions to people on how to provide optimal examples improves teaching behavior. The improvement shows in the reduction of the uncertainty in the estimation of the rewards.

Similarly, the work of Cakmak and Thomaz (2014) investigates human teaching behavior in three classification tasks: faces, animals and gestures.

The authors find that natural teaching is not optimal and hypothesize that because human teaching is largely optimized for human learning, they might not understand the inner working of an artificial learner. To improve the teaching behavior they propose teaching guidance and they show that their system guiding the human how to select teaching examples increases the learning performance of the artificial learner by increasing the accuracy.

The work of Sena, Zhao, and Howard (2018) also addresses the problem of how to provide a set of good quality demonstrations by giving teaching guidance to the human. In this work the authors apply teaching guidance to a task, where the robot has to learn a trajectory from a starting zone to a goal. They furthermore give visual feedback on the learner model after the teaching phase. Their approach improves the teaching efficiency that they determine by the ratio of generalisation performance against the required number of demonstrations by approximately 180%.

While we see that there is upcoming research on human teaching behavior to robots or virtual agents, this research is still sparse and limited to simple tasks. Some of this research focuses on teaching humans how to be better teachers, and the research aiming for understanding human behavior and how to better learn from it is even more sparse. In Chapter 7 we address this by investigating human teaching behavior to a robot for a sensorimotor task.

Chapter 5 Observer Related Metrics

5.1 Introduction

Terms like explicability (Kulkarni et al., 2019; Zhang et al., 2017), legibility (Dragan, Lee, and Srinivasa, 2013), transparency (MacNally et al., 2018; Broekens and Chetouani, 2019) and predictability (Fisac et al., 2018) have become popular in recent research on artificial agents. These terms describe, depending on their definition, similar or contradicting concepts. A comprehensive overview of different concepts is presented in (Chakraborti et al., 2018) and (Wallkotter et al., 2020). All concepts have in common that they assume some kind of observer that tries to infer the intentions of the agent. This idea goes along with the concept of ToM (Section 2.3).

In order to implement these concepts with robots it is important to formalize these kind of concepts. A framework to formalize goal-to-action inference and action-to-goal inference (see Section 2.3) of trajectories is presented by Dragan, Lee, and Srinivasa (2013). Despite predictability and legibility not being the only observer related metrics proposed in literature, here we will present these two in more detail, matching the scope of the thesis.

While predictable (goal-to-action) and legible (action-to-goal) motion trajectories can correlate, they are *"fundamentally different and often contradictory properties of motion"* (Dragan, Lee, and Srinivasa, 2013). While legibility requires the knowledge of possible goals: *"Plan legibility reduces ambiguity over possible goals that might be achieved"* (Chakraborti et al., 2018),

predictability requires the knowledge of a goal/planning problem: *"Plan predictability reduces ambiguity over possible plans, given a goal/planning problem"* (Chakraborti et al., 2018).

Both concepts have in common that some kind of observer of the actions of the robot exists. Consider the example shown in Fig. 5.1 to illustrate the concepts. The left side depicts a predictable movement, if it's known that the robot's gripper is moving for the green (to the right) goal the shown trajectory is the trajectory one would expect. The right side depicts a legible movement. In this case the goal the robot is aiming for is not known beforehand, but if the robot's gripper moves in the shown way, it's very likely early on that it's moving for the green goal (to the right).

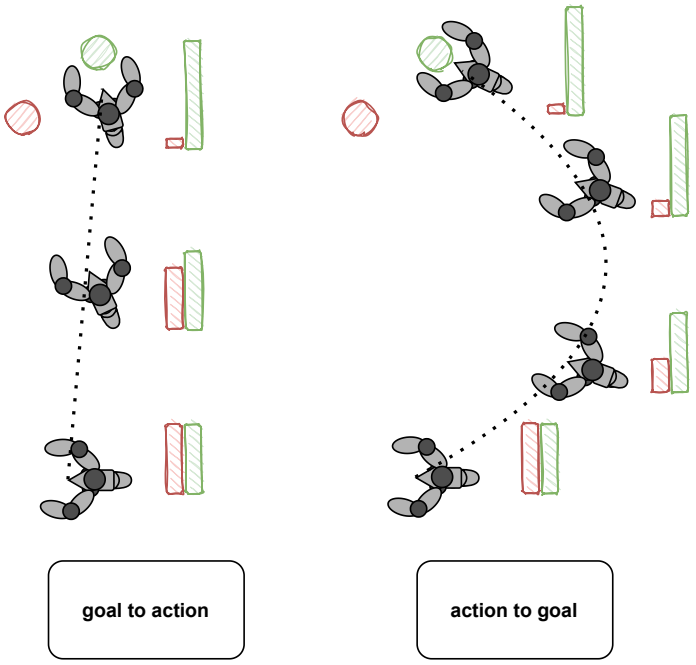


Figure 5.1: Example for a predictable (left) and predictable (right) trajectory for a robot gripper grasping for the green goal. For the legible trajectory (corresponding to action to goal inference) the likelihood that the gripper is going for the green object increases faster than for the predictable (corresponding to goal to action inference) trajectory. Partial recreation from Dragan, Lee, and Srinivasa (2013).

5.2 Legibility

As previously mentioned, legibility implements the idea of action-to-goal inference, thus a legible motion needs to enable the observer to infer the correct goal with high confidence. Here we present the legibility metric as proposed by Dragan, Lee, and Srinivasa (2013). The inference function $\mathcal{I}_L$ maps (snippets of) trajectories from all trajectories Ξ to the set of goals $\mathcal{G}$ and can be denoted as:

$$\mathcal{I}_L = \Xi \rightarrow \mathcal{G} \tag{5.1}$$

The inference needs to happen as fast as possible with high confidence. These properties of legibility are captured by the following equation:

$$\lambda(\xi) = \frac{\int P(g^*|\xi_{s_0 \rightarrow s_t}) f(t) dt}{\int f(t) dt} \tag{5.2}$$

We integrate over the probability to infer the correct goal (the target) given the current trajectory $P(g^*|\xi_{s_0 \rightarrow s_t})$. Therefore, higher inference probability of the target will result in a higher legibility.

The second requirement is that this inference should happen as fast as possible. $f(t)$ provides a simple function to give higher weights to earlier parts of the trajectory, achieving this goal. The following equation provides a simple implementation of $f(t)$:

$$f(t) = T - t \tag{5.3}$$

with T as duration of the trajectory as suggested. As a next step the probability $P(g|\xi_{s_0 \rightarrow q})$ needs to be calculated. This step can be done starting of with Bayes's Rule:

$$P(g|\xi_{s_0 \rightarrow q}) \propto P(\xi_{s_0 \rightarrow q}|g) P(g) \tag{5.4}$$

$P(g|\xi_{s_0 \rightarrow q})$ is the probability that the agent follows $\xi_{s_0 \rightarrow q}$ when the agent targets a possible goal $g \in \mathcal{G}$. q can be any intermediate point. The prior probability of a goal $P(g)$ is assumed to be known, otherwise a uniform prior can be used.

$P(\xi_{s_0 \rightarrow q}|g)$ can be computed as the ratio of all trajectories from s_0 to g that

pass through $\xi_{s_0 \to q}$ to all trajectories from s_0 to g:

$$P(\xi_{s_0 \to q}|g) = \frac{\int_{\xi_{q \to g}} P(\xi_{s_0 \to q \to g})}{\int_{\xi_{s_0 \to g}} P(\xi_{s_0 \to g})} \tag{5.5}$$

Following the assumption that trajectories are separable (Ziebart et al., 2008), i.e. $P(\xi_{s_0 \to q \to g}) = P(\xi_{s_0 \to q})P(\xi_{q \to g})$, leads to:

$$P(\xi_{s_0 \to q}|g) = \frac{P(\xi_{s_0 \to q}) \int_{\xi_{q \to g}} P(\xi_{q \to g})}{\int_{\xi_{s_0 \to g}} P(\xi_{s_0 \to g})} \tag{5.6}$$

At this point, a model is required to express the probability of a trajectory in the eyes of an observer $P(g|\xi_{s_0 \to q})$. The principle of maximum entropy as suggested by (Ziebart et al., 2008) is adopted to model this probability as $P(\xi) \propto \exp(-C(\xi))$.

$C(\xi)$ is the cost associated with trajectory ξ, therefore the probability of a trajectory decreases exponentially with increasing costs, leading to:

$$P(\xi_{s_0 \to q}|g) \propto \frac{\exp(-C(\xi_{s_0 \to q}) \int_{\xi_{q \to g}} \exp(-C(\xi^*_{q \to g}))}{\int_{\xi_{s_0 \to g}} \exp(-C(\xi^*_{S \to G}))} \tag{5.7}$$

These integrals are computationally challenging and Dragan and Srinivasa (2012) derive an approximation with the assumptions that C is quadratic and its Hessian is constant. Under these assumptions according to Laplace's method we have $\int \exp(-C(\xi_{s_0 \to q}))) \approx k \exp(-C(\xi^*_{s_0 \to q}))$, with the constant k and $\xi^*_{s_0 \to q}$ as the optimal trajectory from s to q w.r.t. C.

Using a normalization factor z calculated with:

$$z = \sum_{\mathcal{G}} P(g|\xi_{s_0 \to q}) \tag{5.8}$$

and plugging this expression into Eq. 5.7 leads to:

$$P(g|\xi_{s_0 \to q}) = \frac{1}{z} \frac{\exp(-C(\xi_{s_0 \to q}) - C(\xi^*_{q \to g}))}{\exp(-C(\xi^*_{s_0 \to g}))} P(g) \tag{5.9}$$

Approximating the cost C with the quadratic trajectory length in workspace

punishes the agent from unnecessarily long paths:

$$C = \sum_t \|\xi_{s_0 \to s_{t+1}} - \xi_{s_0 \to s_t}\|^2 \tag{5.10}$$

Using C as stated in Eq. 5.10 can only serve as rough approximation, since knowing the real cost function the observer will associate with the movement is a major challenge.

In situations with multiple goals, an agent can make trajectories more and more legible and never reaching a score of one while increasing the cost w.r.t to C more and more. In order to prevent the agent to go to far away from the observer's expectation, a regularizer $L(\xi)$:

$$L(\xi) = \lambda(\xi) - \mu C(\xi) \tag{5.11}$$

can be used.

The legibility metric was not only derived theoretically, but Dragan, Lee, and Srinivasa (2013) show furthermore in an experiment with real humans that for legible trajectories the participants were faster able to infer the target goal with higher probability correctly.

Dragan and Srinivasa (2013) extend the work and use the framework to generate legible (motion) trajectories by introducing constrained legibility optimization. This framework is also used in Holladay, Dragan, and Srinivasa (2014) to create legible pointing trajectories.

5.3 Predictability

Predictability captures the idea of the goal-to-action inference. In this case the goal is known a priori. Before the robot moves the observer will create an expectation how the robot will move, thus infer a trajectory given the goal.

The more the robot moves like the observer expects, the more predictable the trajectory is. This inference can be denoted as a mapping from goals to

trajectories as follows:

$$\mathcal{I}_P : \mathcal{G} \to \Xi$$

Predictable motion is then formalized as motion for which the trajectory $\xi_{S \to G}$ matches this inference:

$$\mathcal{I}_P(G) = \xi_{S \to G}$$

It seems like a reasonable assumption that agents (humans or agents) will try to minimize their costs while executing an action, thus the most predictable trajectory is the one associated with the lowest cost C:

$$\mathcal{I}_P = \arg \min_{\xi \in \Xi_{S \to G}} C(\xi) \tag{5.12}$$

A predictability score that is normalized from 0 to 1 can then be calculated with:

$$\text{predictability}(\xi) = \exp(-C(\xi)) \tag{5.13}$$

Thus maximizing this score is equivalent to minimizing the cost function. As already mentioned in Section 5.2, knowing the real cost function that is assumed by the observer is difficult, and depending on the actual function the minimization can also be challenging.

While we do not use predictability, we use legibility and the goal probability as presented here in Chapter 8 to model observer feedback to an RL agent and legibility as proxy how well a potential observer might reason about the goals of the agent.

Part III

Implementation of Research

Chapter

6 Communication Model

Contents

6.1 Introduction

In this chapter we start by proposing a general communication model for HRI settings. The model builds on common ideas found in HRI research, and as such does not provide novel ideas. However, explicitly introducing a model is useful to position research that focus on the channel usage in HRI settings. While a similar model has been proposed by Sigaud et al., 2021 (see Section 2.4), the model does not take signals we are mainly interested in into account. These signals are namely signals that combine social and task signals in one signal. After introducing the general model, we present our specific approach focusing on only one channel that combines social and task signals, while not having any other channels. Furthermore, we present how the model relates to our research questions and the implemented research presented in later chapters.

6.2 General Communication Model

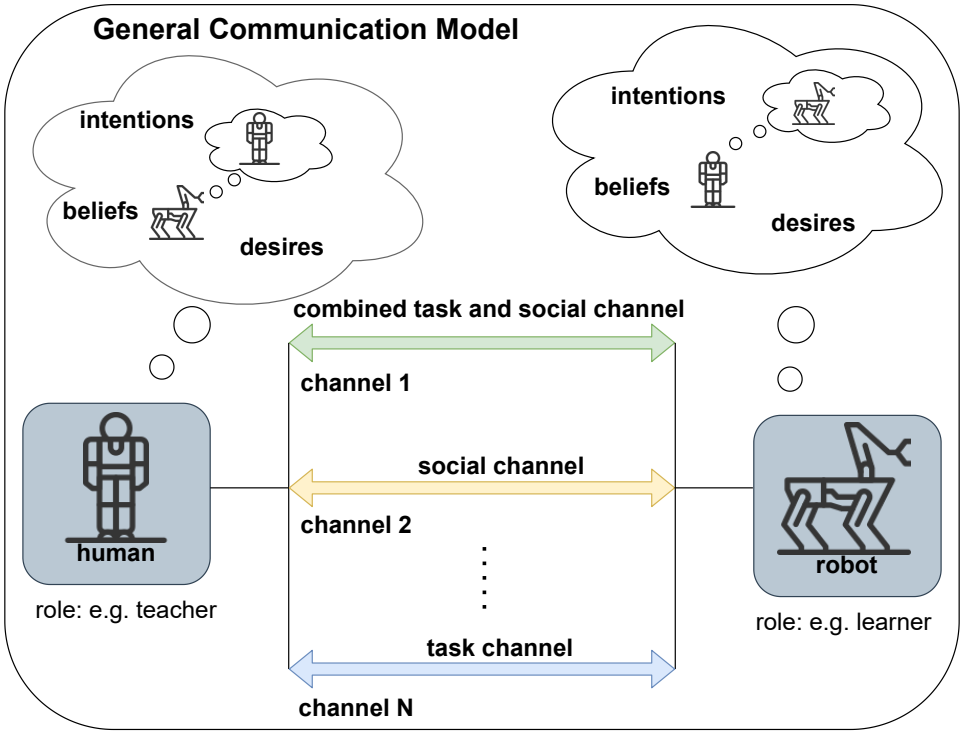


Figure 6.1: Proposed general communication model. The communication happens over any number of channels N. These channels can either be task channel, social channel or combined channel. The communication partner mutually reason about the mind state of the other partner.

Based on related research as presented in Section 2.4 we propose a model to conceptualize communication in HRI settings as shown in Fig. 6.1. While the general model is useful to describe the general setting, in this thesis we focus on communication that combines social and task oriented information in one signal (e.g. SMC Section 2.6).

An important key idea of the model is that the interaction can have social- as well as task aspects (Castellano, Pereira, et al., 2009; Castellano, Leite, and Paiva, 2017; Leclère et al., 2016; Ivaldi et al., 2014) (see Section 2.4). This idea is represented by having a social- and a task channel, as already present in the model proposed by Sigaud et al. (2021). In our general model the communication can happen over any (probably small) number of channels. The channels can either be a task channel, social channel or a combined task and social channel. The combined channel is used to transmit

signals that combine social information as well as information on the task, thus it is important that the receiver recognizes the communicative intent of the sender (see ostensive-inferential communication Section 2.5).

The channel usage is influenced by ToM (see Section 2.3). Both parties reason about the behavior of the communication partner and assume that the communication partner reasons about them too. The beliefs, intentions and desires of the communication partners influence the use of the communication channel.

One important influence factor on the model is if the communication happens in a pedagogical situation (see Section 4.2) or not. In a pedagogical situation one communication partner will have the role of a teacher and the other communication partner the role of the learner. The teacher is the communication partner that selects information to help the learner acquire new knowledge.

6.3 Specific Approach

6.3.1 Specific Model

In our specific approach we limit ourselves to only only one channel that combines task and social signals as shown in Fig. 1.1. As research on SMC shows (see Section 2.6) people take advantage of task- and social channel properties when using the sensorimotor channel. Thus, SMC is good representative for such a channel.

Furthermore, we distinguish between different types of goals, actions and intentions as defined in the following. While we define our own terminology, similar (yet slightly different) definitions can be found in Ho, Cushman, et al. (2021) and Shafto, Goodman, and Frank (2012).

The objective of a communicative goals is that the interaction partner has certain information or knowledge. Communicative actions aim to communicative information to an interaction partner. Communicative intentions link communicative goals to communicative actions.

Instrumental goals have a different state of the environment as an objective. Instrumental actions aim to influence the state of the environment. Instrumental intentions link instrumental goals to instrumental actions. Instrumental actions

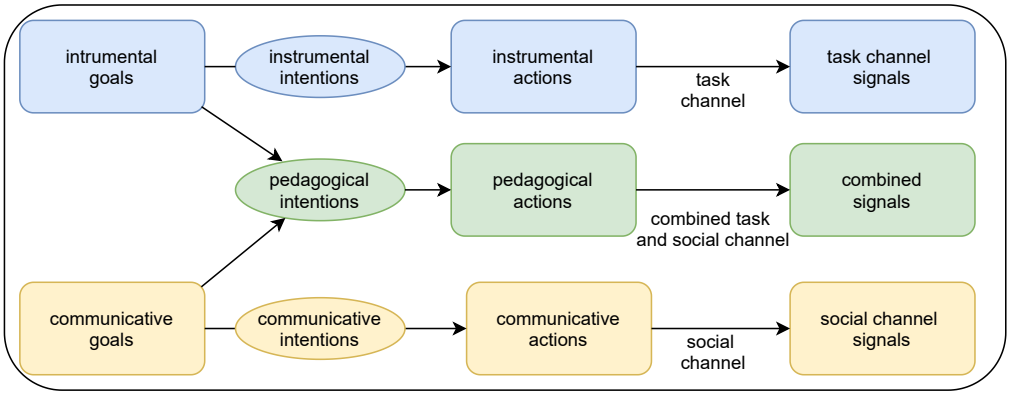


Figure 6.2: Different types of goals are linked to the corresponding actions via the corresponding types of intentions. Depending on the type of action, the actions provide different types of signals.

provide task channel signals when observed by an observer.

Pedagogical intentions combine instrumental goals and communicative goals and link them to pedagogical actions. Pedagogical actions serve two purposes: they modify the environment to achieve task goals and communicate additional information to the interaction partner. The relation between the different types of goals, actions and intentions is illustrated in Fig. 6.2.

The specific model we propose for a pedagogical situation where the teacher only uses one combined task and social channel to communicate to the learner is shown in Fig. 6.3. The teacher has instrumental goals w.r.t. the state of the environment and communicative goals w.r.t. to a certain hypothesis they want to communicate to the learner. The pedagogical intentions of the teacher link the instrumental- and communicative goals to pedagogical actions. The teacher reasons about the state of mind of the learner and chooses pedagogical actions accordingly. The pedagogical actions modify the environment and provide task signals as well as social signals to the learner. The learner reasons about the state of mind of the teacher and infers from the pedagogical actions the (ideally true) hypothesis. The learner can then communicate task- and social signals back. Note that we depict the back channel from learner to teacher in a simple manner, however in general this communication could make use of rich communication features.

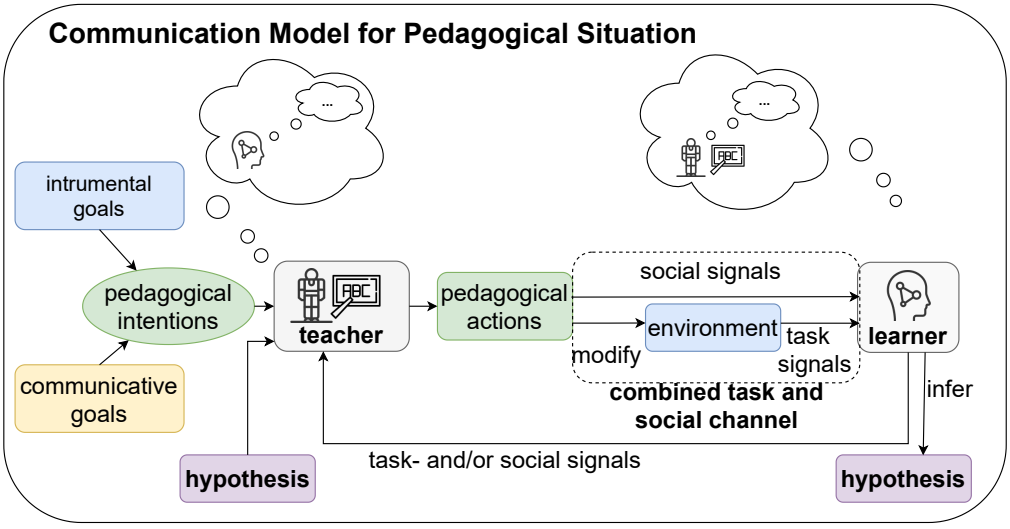


Figure 6.3: Model for communication in a pedagogical situation. The teacher (either robot or human) has instrumental goals to bring the environment into a certain state, as well as communicative goals to communicate certain hypothesis to the learner. The pedagogical intentions of the teacher link both type of goals to pedagogical actions. These actions provide task signals by modifying the environment and provide social signals to the learner at the same time. The learner uses the signals to infer the hypothesis and gives feedback to the teacher.

6.3.2 Model Application to Implemented Research

In the previous section we described our full specific model having a full interaction loop. In this section we describe how the specific model relates to our research questions and how we (partially) apply it in our implemented research.

Application to the User Study

In our user study (see Chapter 7), we focus on the human side of the communication by addressing the questions *Do humans make use of social channel characteristics when teaching robots a sensorimotor task?* (Q1) and *Are negative demonstrations useful to enrich approaches that use demonstrations to learn?* (Q2).

In the first experiment of the user study (see Chapter 7) we address the question *Does human behavior change when teaching a robot how to solve a task as opposed to just solving the task?* (Q1a). We do this by comparing

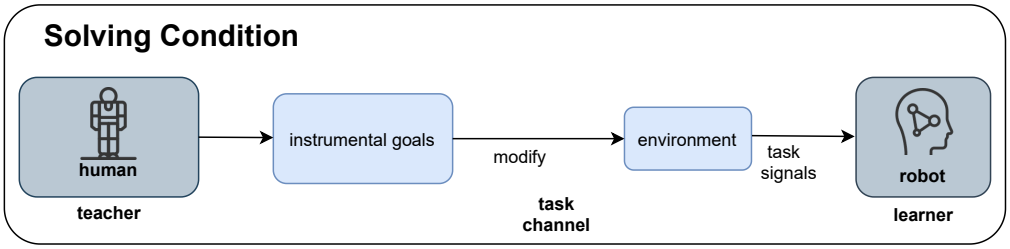


Figure 6.4: Specific model for the solving condition in our user study (see Chapter 7) according to our hypothesis. The robot learns from the task signals, but the human is not an explicit teacher and does not have pedagogical intentions. Thus, the human will only modify the environment providing only task signals to the learner.

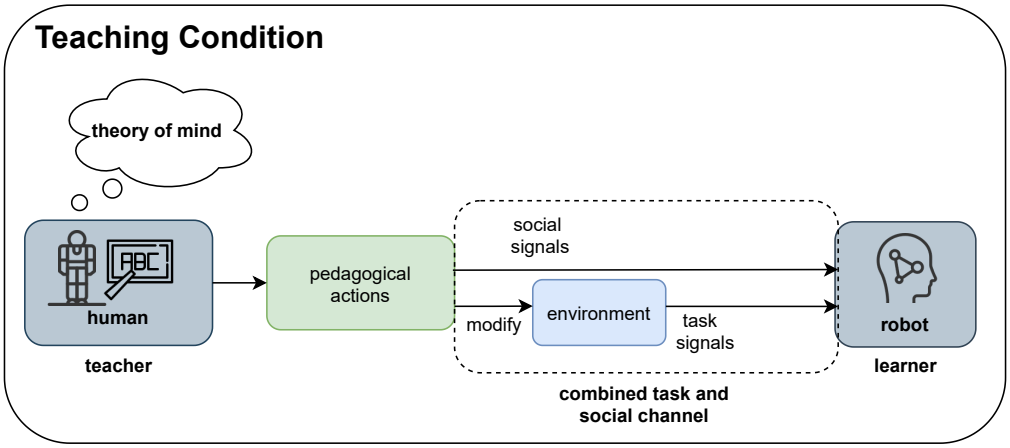


Figure 6.5: Specific model for the teaching condition in our user study (see Chapter 7) according to our hypothesis. The human is the (explicit) teacher and the robot is the learner. The human human will provide task signals by modifying the environment and additional social signals to the learner robot.

two conditions with each other: In the first condition we ask humans solve a sensorimotor task, and in the second condition we ask humans to teach how to solve a sensorimotor task to the robot.

According to our hypothesis, in the solving condition, humans will use sensorimotor actions just to modify the environment and will not try to communicate anything else to the learner (here the robot). The communication model for the solving condition is shown in Fig. 6.4.

In the teaching condition humans will make use of task- as well of social channel characteristics. The specific model for the teaching condition is shown in Fig. 6.5. In order to address Q2, the teaching condition included negative

examples as well.

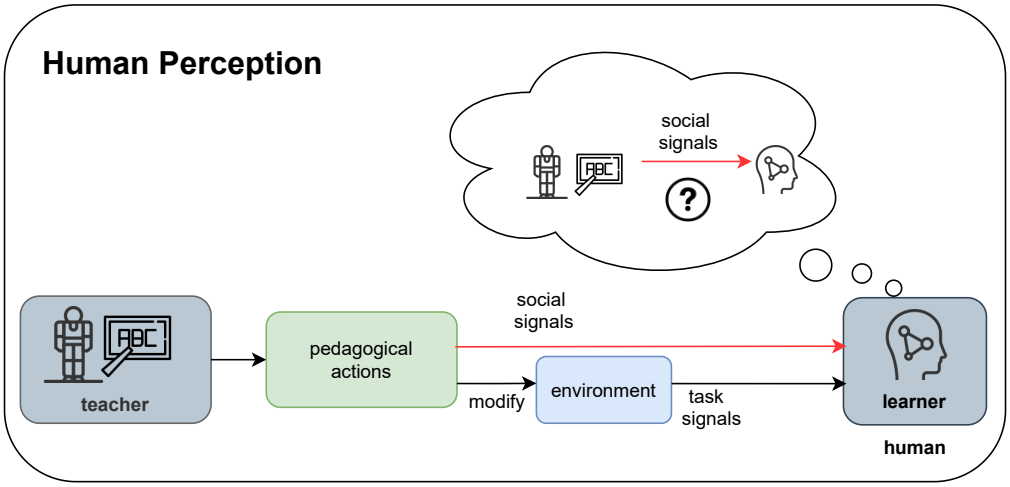


Figure 6.6: Specific model when investigating human perception. The humans does not know from which condition (solving or teaching) the examples came from. The perception on informativeness is used to decide if social signals are present in the actions or not.

In the second experiment of our user study (see Chapter 7) we address the question Do humans perceive this teaching behavior as more informative than the the solving behavior? (Q1b). We do this by showing the demonstrations we collected in the first experiment to new participants. In this setting the human becomes the learner, but does not know from which condition the data came from. However, the teacher here is not a robot, since the participants knew that they were shown data created by humans. The model for the human perception is shown in Fig. 6.6.

Application to the Simulated Experiment

In our third experiment (see Chapter 8) we address the question *How can we integrate actions that make use of social channel characteristics into RL?* (Q3). While the in this experiment the robot learns autonomously without a real human present. The robot has the role of the teacher, the learner is a simulated observer.

During the learning process the robot aims to maximize the goal inference of the observer. The social signal corresponds to the part of the action that aims to maximize the goal inference of the observer. The observer gives feedback about its goal inference to the robot that is integrated into the learning process.

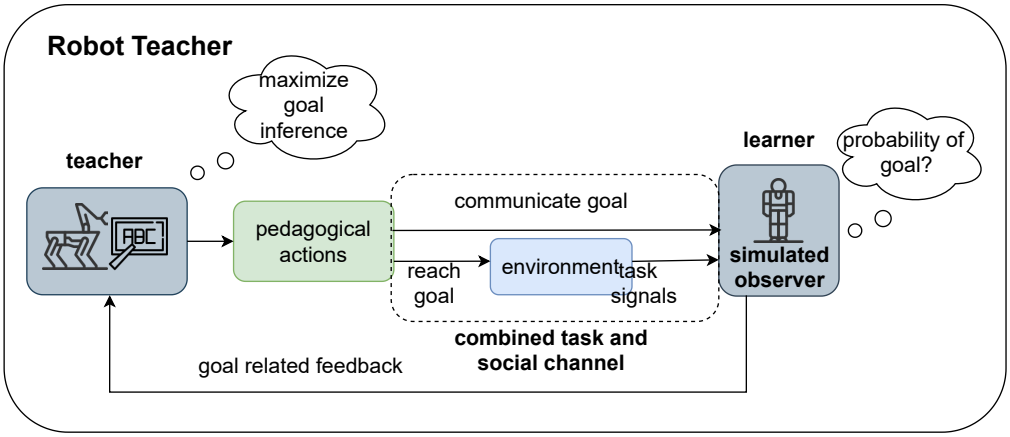


Figure 6.7: Specific model with robot as the teacher in our RL framework (see Chapter 8). When we focus on the robot communicating additional information, the robot becomes the teacher and the human the learner. The robot uses the sensorimotor channel as task- and as social channel.

While we do not use a full ToM, the we use the legibility metric (Section 5.2) as a simplified approximation how well the observer understands the (goal) intentions of the robot. The corresponding model for this experiment is shown in Fig. 6.7.

Chapter 7 User Study on Human Teaching Behavior Towards Robots in a Sensorimotor Task

Contents

7.1 Introduction

LfD is a popular approach to transfer new skills to an agent (e.g. robot) in an intuitive manner (Argall et al., 2009; Billard et al., 2008) (see Section 3.5). LfD research usually assumes an expert giving correct demonstrations, or novice user giving (possibly) flawed demonstrations. The common assumption is that the user just solves the task, giving her best possible solution. However, we could also imagine that the user includes additional information in a demonstration, than simply solving the task alone. Following the *pedagogy hypothesis* (see Section 4.2) this assumption seems reasonable. A context like this corresponds to a pedagogical situation (see Section 4.2).

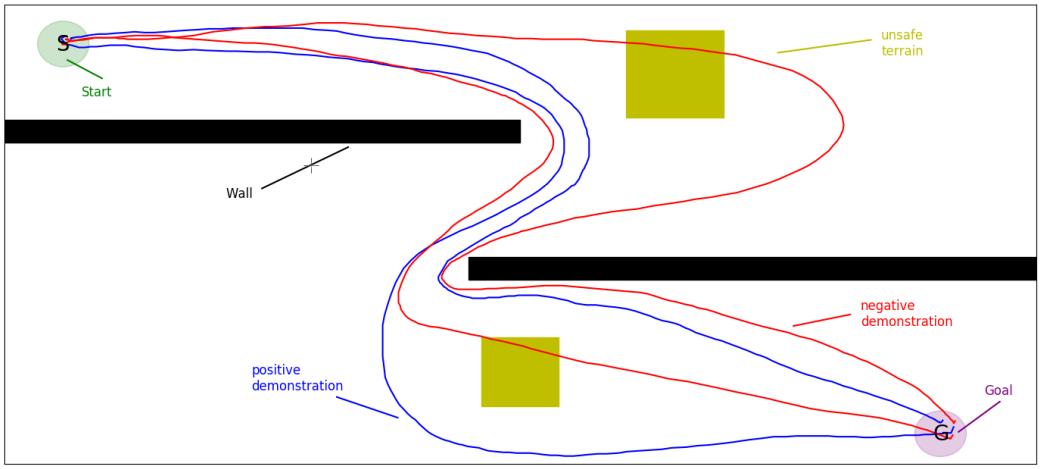


Figure 7.1: Environment of the continuous maze task. Humans have to solve the continuous maze task (Solving) or teach how to solve it (Teaching) with the possibility to provide negative demonstrations.

The interesting question here is whether people change their demonstrations in these pedagogical situations, and if so, how do they change such demonstrations? The work of Ho, Littman, MacGlashan, et al. (2016) shows that there is a difference between people solving and showing how to solve a 2-dimensional grid world task Section 4.2. In the context of physical interaction, SMC (Pezzulo, Donnarumma, and Dindo, 2013; Pezzulo, Donnarumma, Dindo, et al., 2019) is closely related to this question (see Section 2.6).

Calinon (2019) identifies the exploitation of the social interaction as future direction of LfD research, and gives learning from counterexamples as example. Further, the work of Osa et al. (2018) identifies the learning from different instruction types as a challenge for LfD research.

Similarly to counterexamples, we propose to allow teachers to provide negative demonstrations as one approach to address this challenge. Negative demonstrations are demonstrations that explicitly demonstrate what not to do. Related research (e.g. Breazeal, Berlin, et al., 2006; Mueller, Venicx, and Hayes, 2018; Cui and Niekum, 2018) integrates possibilities to learn from sub-optimal and flawed demonstrations, but does not offer the possibility to purposefully demonstrate what not to do.

In our study we are interested in how humans use sensorimotor communication to teach a sensorimotor task. We are interested in the the difference between

solving and teaching, furthermore we explore how humans use the possibility of using negative demonstrations to teach the task. Further, by varying the number of demonstrations people could give we introduce a notion of costs for a demonstration related to the ideas of MT (see Section 4.3).

We had three hypotheses:

1. Humans modify their behavior when teaching how to solve a sensorimotor task in comparison to solving it.
2. Humans perceive the use of negative demonstrations as informative.
3. The teaching behavior depends on the number of demonstrations people are allowed to give.

This chapter relates to the bigger picture of this thesis by focusing on the human side of the communication and address the questions *Do humans make use of social channel characteristics when teaching robots a sensorimotor task?* (Q1) and *Are negative demonstrations useful to enrich approaches that use demonstrations to learn?* (Q2).

The results presented in this chapter have (partially) been published in Bied and Chetouani (2020a).

7.2 Study

7.2.1 Overview

We designed and conducted a study to investigate the difference between humans solving and teaching a task using the sensorimotor channel. The study was conducted at the INSEAD-Sorbonne Université Behavioural Lab (INSEAD). Further, it was reviewed by INSEAD's Institutional Review Board (IRB) under the reference number 201913. The IRB approval letter is attached in Appendix A.1.

The texts of the study were designed in English and translated by INSEAD in dialogue with us to French. Both parts of the study were conducted in French. Also the participants were recruited by INSEAD. We briefed and debriefed the participants before and after each of the two experiments. Further, the participants were thanked and compensated for their time according to standard

rates. The consent- and debriefing document for experiment 1 are attached in Appendix A.2 and Appendix A.3, the consent- and debriefing document for experiment 2 are attached in Appendix A.4 and Appendix A.5-A.6.

The study consisted of two experiments:

1. We asked participants first to solve a sensorimotor task, and afterwards to teach how to solve it to a robot.
2. We asked new participants to rate the demonstrations from the first experiment.

Our specific communication model corresponding to the solving condition is shown in Fig. 6.4. We hypothesize that in the solving condition, humans will use sensorimotor actions just to modify the environment and will not try to communicate additional information to the robot.

Our specific communication model for the teaching condition is shown in Fig. 6.5. We hypothesize that in the teaching condition humans will make use of the combined task and social channel by modifying the environment and including additional social signals. Negative demonstrations are included to address Q2.

Furthermore, we are interested in human perception to address the question *Do humans perceive this teaching behavior as more informative than the the solving behavior?* (Q1b). We address this by asking questions to the participants about their perception. Further, by asking the participants of experiment 2 about their perception on the collected examples from experiment 1. Our corresponding model for the human perception is shown in Fig. 6.6.

7.2.2 Experiment 1

Description

The first experiment was conducted with 42 participants (21 female, 21 male, Ø of age = 23.76 years, σ^2 of age = 18.65). We asked the participants to teach how to solve a continuous maze task to a robot. In order to add the sensorimotor dimension the task was solved on tablet with a digital pen.

The task consists of going from a start zone to a goal zone. The environment has impassable terrain in black, and unsafe terrain in yellow. The impassable

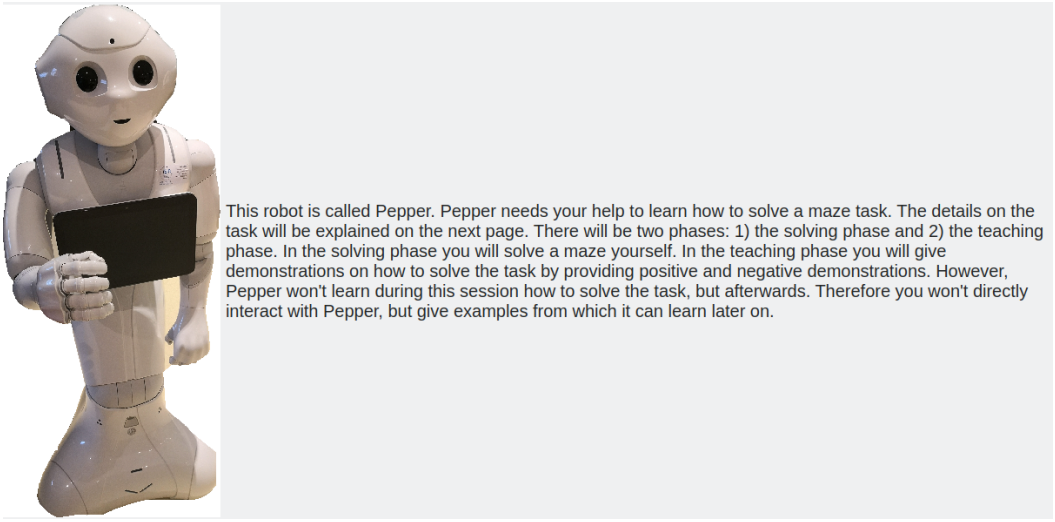


Figure 7.2: Introduction of the pepper robot to the participants. They were explained that they would not directly interact with the robot, but would give demonstrations that the robot could later use to learn how to solve the task.

terrain can not be crossed. The unsafe terrain could be crossed, but is as the name states; unsafe. An instance of the maze task can be seen in Fig. 7.1.

The experiment includes a *Solving*- and a *Teaching*-phase, in order to show that there is a difference between solving and teaching (similar to Ho, Littman, MacGlashan, et al., 2016). In the Solving-phase, the participants were just asked to solve the task correctly. In the Teaching-phase, the participants were asked to give positive and negative demonstrations. Positive demonstrations are correct solutions to the task. Negative demonstrations go through the unsafe zone, but were also required to start at the goal zone and end in the end zone. The Solving- and Teaching-phase was repeated for 15 different instances of the task in an alternating manner. The environments used in the experiment are attached in Appendix B.

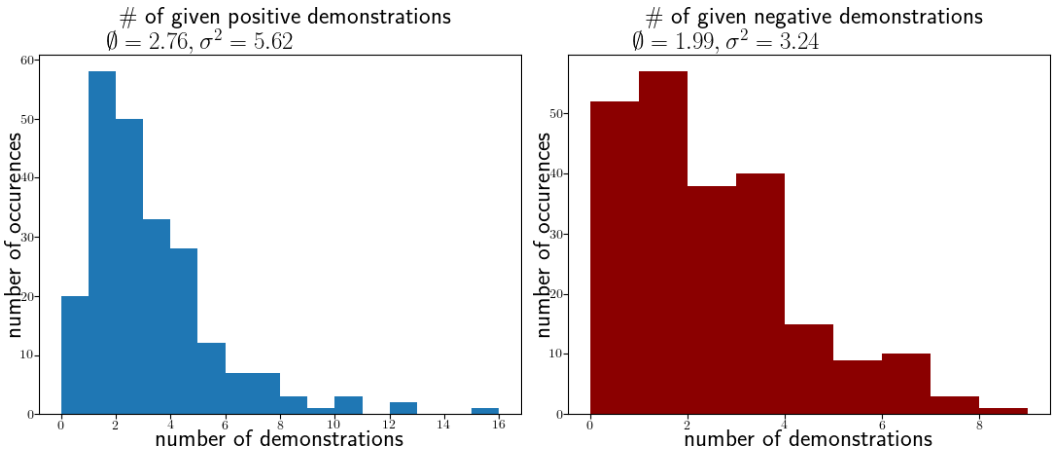
The participants where randomly split into three conditions that differed in the number of demonstrations they were allowed to give in the Teaching-phase. In the first condition they were allowed to give any number of demonstrations, in the second condition they could only give one demonstrations, and in the third condition they were asked to give three demonstrations. The conditions did not differ in the Solving-phase where the participants were asked to solve the task exactly once. These different conditions were implemented to introduce a notion of cost with MT (see Section 4.3) in mind.

The participants did not directly interact with a robot, but were introduced to the pepper robot with a picture and a description as shown in Fig. 7.2. Additionally, they were told that the robot would learn from their demonstrations later.

After completing the Solving- and Teaching phase for all environments, the participants were asked to answer the following four questions on a 5-point-Likert scale:

1. How useful are negative demonstrations?
2. How difficult was it to give negative demonstrations?
3. How useful are positive demonstrations?
4. How difficult was it to give positive demonstrations?

Results



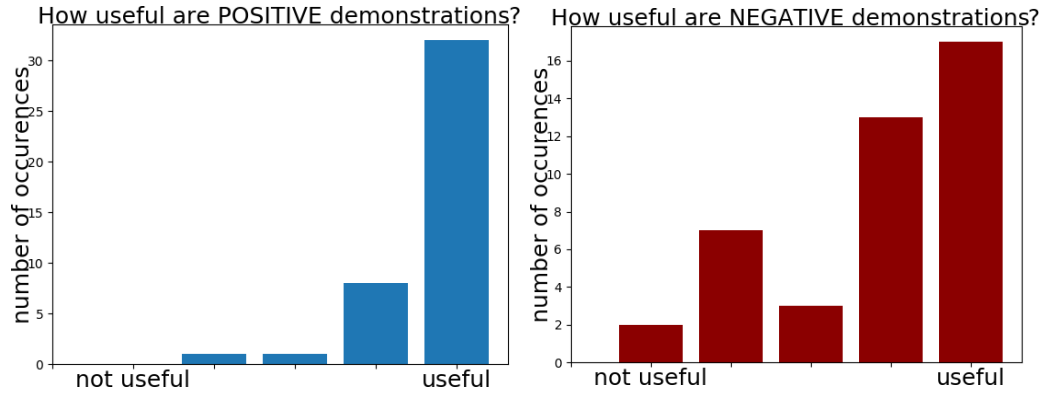
(a) Number of occurrences for positive demonstrations.

(b) Number of occurrences for negative demonstrations.

Figure 7.3: Number of occurrences for different counts of positive and negative demonstrations given in the first condition.

The participants in the first condition could give any number of demonstrations in the Teaching-phase. On average, the participants gave 2.76 ($\sigma^2 = 5.62$) positive demonstrations and 1.99 ($\sigma^2 = 3.24$) negative demonstrations for each environment. Thus, on average, the number of given demonstrations is higher for positive- than for negative demonstrations, while also the number of positive demonstrations varies more for positive demonstrations. The bar plot showing

the number of occurrences plotted against the number of given demonstrations in a singular demonstration is shown in Fig. 7.3.

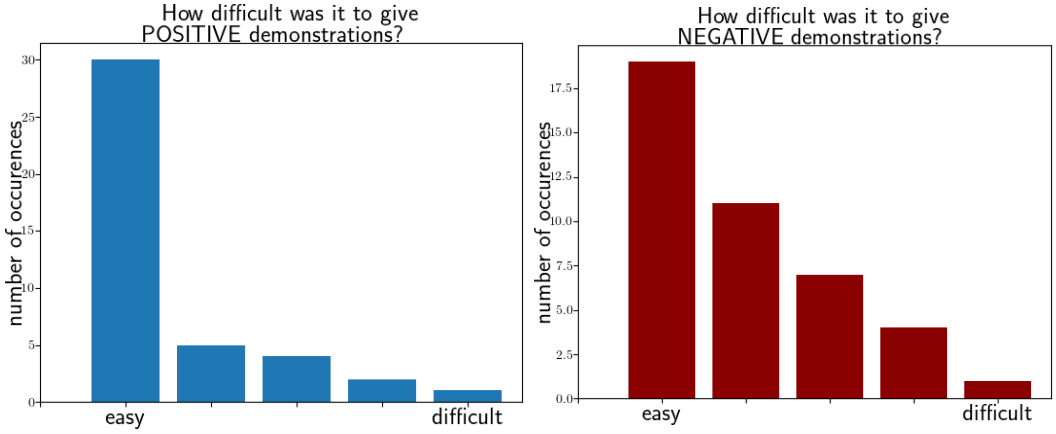


(a) Answers on the usefulness of positive demonstrations.

(b) Answers on the usefulness of negative demonstrations.

Figure 7.4: Occurrences of the answers (on a likert-scale) for the question how useful negative, respectively positive demonstrations are.

The results for the question on the usefulness of negative- respectively positive demonstrations are shown in Fig. 7.4. A majority of participants consider positive demonstrations as well as negative demonstrations as useful. However, the picture is clearer for the positive demonstrations. While 95% of the participants consider positive demonstrations as useful, only 71% of the participants consider negative demonstrations as useful.



(a) Answers on the difficulty of negative demonstrations.

(b) Answers on the difficulty of positive demonstrations.

Figure 7.5: Occurrences of the answers (on a likert-scale) for the question how difficult it is to give negative, respectively positive demonstrations.

The results for the question on the difficulty of negative- respectively positive demonstrations are shown in Fig. 7.5. Similarly to the results on the previous questions, a majority of participants considers giving positive demonstrations as well as negative demonstrations as easy. The tendency here is also stronger for the positive demonstrations. While 83% of participants consider giving positive demonstrations as easy, only 71% of participants consider giving negative demonstrations as easy.

7.2.3 Experiment 2

Description

In the second experiment we asked 72 (38 female, 32 male, 2 none of the before mentioned, Ø of age = 23.65 years, σ^2 of age = 12.23) new participants to rate the collected demonstrations, discarding the demonstrations that did not fulfill the requirements of a correct positive, respectively negative demonstration. A condition for the recruitment for this experiment was that the participants did not already participate in the previous experiment.

The participants received a description of the maze task from the previous task. Then they were explained how the teaching process works. Afterwards, they were asked to rate the statement: *"The demonstrator deviates from the simplest way of solving to convey to you other information about the task"* on a 5-point-Likert scale from *strongly disagree* to *strongly agree*. Each demonstration was rated by at least 6 participants. The Graphical User Interface (GUI) used for the experiment is shown in Appendix C.1-C.3.

Results

The ratings clustered, whereas ratings of 1 and 2 were counted as *Non-informative*, ratings of 3 as neutral and ratings of 4 and 5 as *Informative*. Each demonstration was classified according to a majority voting between all participants that gave a rating on a particular demonstration, demonstrations that did not have a majority for neither Non-informative nor Informative were counted as *Not-clear*.

The absolute numbers of the classified demonstrations are reported in Table 7.1. In the Solving-phase there were no negative demonstrations possible (N/A),

Table 7.1: Absolute values of the results of the majority votes how informative an example is.

		Non-informative	Not-clear	Informative
positive demonstrations	Solving	499	48	44
	Teaching	851	132	206
negative demonstrations	Solving	N/A	N/A	N/A
	Teaching	198	174	513

Table 7.2: Relative numbers for Solving (in %).

	Non-informative	Not-clear	Informative
positive demonstrations	84.43	8.12	7.45

the relative numbers for the Solving-phase can be seen in Table 7.2 and for the Teaching-phase in Table 7.3. In the set of positive demonstrations, the relative portion of informative demonstrations in the Teaching-phase (17.33%) is higher than in the Solving-phase (7.43%).

This difference between the Solving- and Teaching-phase is significant considering the columns for Non-informative, Not clear and Informative, $\chi^2(2, N = 1780) = 39.52$, $p<.01$, as well when only considering the columns for Non-informative and Informative, $\chi^2(1, N = 1600) = 34.42$, $p<0.01$.

In the Teaching-phase, the relative portion of informative demonstrations is significantly higher for the negative demonstrations (57.97%) than for the positive demonstrations (17.33%). This difference between positive and negative demonstrations is significant considering the rows for Non-informative, Not-clear and Informative, $\chi^2(2, N = 2074) = 509.73$, $p<0.01$, as well when only using the row for Non-informative and Informative, $\chi^2(1, N = 1768) = 486.39$, $p<0.01$.

Table 7.4 shows the absolute numbers divided after the three conditions for

Table 7.3: Relative numbers for Teaching (in %).

	Non-informative	Not-clear	Informative
all demonstrations	50.58	14.75	34.67
positive demonstrations	71.57	11.1	17.33
negative demonstrations	22.37	19.66	57.97

Table 7.4: Results of the classification divided after conditions for positive demonstrations (Teaching) in absolute numbers.

# of demonstrations	Non-informative	Not-clear	Informative
any (Ø=2.76, $\sigma^2 = 5.62$)	529	88	130
1	329	19	25
3	492	73	95

Table 7.5: Results of the classification divided after conditions for positive demonstrations (Teaching) in relative numbers (in %).

# of demonstrations	Non-informative	Not-clear	Informative
any (Ø=2.76, $\sigma^2 = 5.62$)	70.82	11.78	17.40
1	88.2	5.09	6.70
3	74.55	11.06	14.39

the positive demonstrations, the corresponding relative numbers are shown in Table 7.5. We see that the portion of Informative demonstrations is with 6.7% the lowest for condition 2 (1 demonstration), with 14.39% a little higher for condition 3 (3 demonstrations) and with 17.40% the highest for condition 1 (any number of demonstrations). This difference is significant when considering all columns ($\chi^2(4, N = 1780) = 42.545$, p<0.01), as well when omitting the Not-clear column ($\chi^2(1, N = 1600) = 28.238$, p<0.01).

Table 7.6 shows the absolute numbers divided after the three conditions for the negative demonstrations, the corresponding relative numbers are shown in Table 7.7. We see that the portion of Informative demonstrations in condition 2 (1 demonstration) is with 53.79% only a little lower than for condition 3 (3 demonstrations) with 58.33% and condition 1 (any number of demonstrations) with 59.24%. This difference is neither significant when considering all columns ($\chi^2(4, N = 885) = 1.338$, p>0.05) nor significant when omitting the Not-clear

Table 7.6: Results of the classification divided after conditions for negative demonstrations (Teaching) in absolute numbers.

# of demonstrations	Non-informative	Not-clear	Informative
any (Ø=1.99, $\sigma^2 = 3.24$)	80	70	218
1	35	32	78
3	83	72	217

Table 7.7: Results of the classification divided after conditions for negative demonstrations (Teaching) in relative numbers (in %).

# of demonstrations	Non-informative	Not-clear	Informative
any (Ø=1.99, $\sigma^2 = 3.24$)	21.74	19.02	59.24
1	24.14	22.07	53.79
3	22.31	19.35	58.33

column ($\chi^2(2, N = 711) = 0.704$, p>0.05).

Condition 1 (any number of demonstrations) and condition 3 (3 demonstrations) look quite similar. When omitting the row of condition 2 (1 demonstration), the difference between condition 1 and condition 3 is neither significant for the positive demonstrations ($\chi^2(2, N = 1407) = 2.814$, p>.05), nor significant for the negative demonstrations ($\chi^2(2, N = 740) = 0.064$, p>.05).

7.3 Conclusion

In this study we showed that there is a difference between people solving and teaching a sensorimotor task to a robot confirming our first hypothesis. However, even when the difference between the Teaching- and Showing-phases is significant, for the positive demonstrations only a relatively small portion (7.43% and 17.33%) are in the Informative category, making it difficult to predict from which phase a single demonstration was taken.

The novelty of this work is that we show that people perceive a significant higher portion of negative than positive demonstrations as informative. Further, 58% of the negative demonstrations are informative, indicating that our second hypothesis is also verified. Our third hypothesis, that the teaching behavior depends on the number of demonstrations people are allowed to give is partially confirmed for positive demonstrations, but not for negative demonstrations. One reason could be that, since negative demonstrations are already perceived as rather informative, people do not see the need to increase the informativeness further.

Since we were able to verify our first two hypotheses we conclude that the specific models presented in Chapter 6 for this study are accurate to conceptualize the communication process.

One possibility to extend this work is to train classifiers to automatically detect between Teaching and Solving or training generative models to solve the task. This work also shows that aiming for integrating negative demonstrations into a LfD framework is promising. Another promising future direction shown by this work is the integration of negative demonstrations. Another direction extends into a direction of predictability and legibility (Dragan, Lee, and Srinivasa, 2013) (see Chapter 5).

Chapter 8 Augmenting RL with Social Channel Usage

Contents

8.1 Introduction

Humans and robots working together - so called Human-Robot cooperation - has recently become a popular area of research. This cooperation can allow robots and humans to accomplish more sophisticated tasks. When it comes to cooperation, one crucial difference between humans and other species is

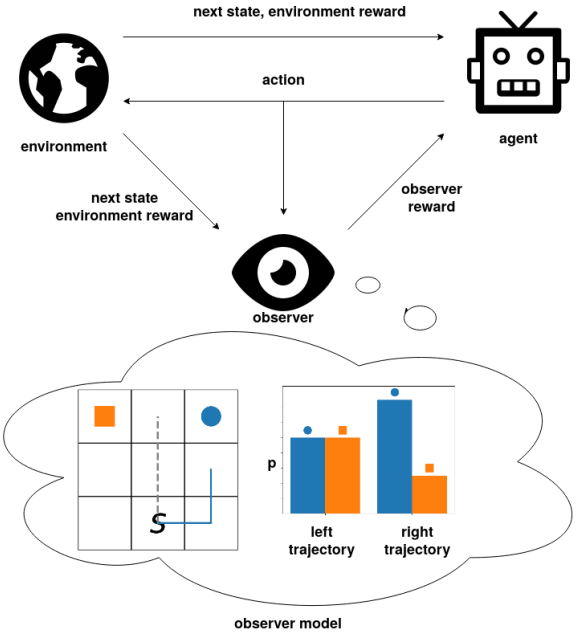


Figure 8.1: Setup of the observer RL framework. As in a regular RL setting the agent interacts with the environment and receives a reward after each taken action. The observer gives an additional reward as feedback to the agent based on how well the observer can infer which of multiple possible goals the agent is targeting. This additional feedback results in the agent learning more legible trajectories.

the capability to share goals and intentions (Tomasello et al., 2005) (see Section 2.3).

In order to mimic these capabilities in Human-Robot collaboration, it is necessary to equip the robot with the capability of shared goals and intentions. One important aspect to achieve this intention sharing is that the robot understands what the human is doing, for example to predict human motions (Mainprice, Hayne, and Berenson, 2015).

Another important aspect is to enrich the robot with behavior that can be well understood by humans. In general, this problem requires the robot to behave more transparently or with explainability (see Chapter 5). Depending on the task, this constraint on the robot's behavior requires the robot's motion trajectories to be either predictable or legible.

This observer reasons about the possible intentions, either possible goals or plans, of the robot. Aforementioned concepts are rarely studied in combination with machine learning. A suitable candidate to use machine learning in this

context is RL Section 3.4.

The classical RL approach does not offer the possibility to add a human to the loop, depriving the framework of integrating valuable task knowledge from the human. This gap has already been addressed in research on interactive RL investigating different models of human-feedback (Thomaz and Breazeal, 2008; Knox and Stone, 2012b; Griffith et al., 2013; MacGlashan et al., 2017; Celemin, Ruiz-del-Solar, and Kober, 2019) (see Section 3.4).

Usually, these approaches have the goal to speed up the learning process of the agent or enable it to find better solutions. To the best of our knowledge, none of the work on interactive RL explores the role of an observer that reasons about the intent of the agent in an interactive RL framework. However, integrating such an observer into interactive RL is an important step towards using RL for human-robot collaboration.

Some similar ideas to our approach can be found in Qi and S. Zhu (2018), where the agents in a multi-agent RL setting integrate the intent of the other agents when calculating the optimal action, but do not express their intent themselves.

Further, Huang et al. (2017) employs different approximate-inference IRL variations to model how humans infer an agent's objective function and use an AT approach (see Section 4.3) to generate a set of environments to increase the probability of inferring the correct objective function. For each environment the optimal trajectory according to the objective function is shown, therefore it's not about comparing different trajectories in one environment like in our approach.

Pezzulo, Donnarumma, and Dindo (2013) proposes to use a signaling distribution of a trajectory in order to facilitate its recognition by another person (see Section 2.6).

The work of Ho, Littman, MacGlashan, et al. (2016) and Ho, Littman, Cushman, et al. (2018) combines the idea of IRL and communication via the means of pedagogical reasoning (Shafto, Goodman, and Griffiths, 2014) (see Section 4.2). However, no other work uses an interactive RL scheme to learn more legible behavior.

Thus, in this chapter we explore how to integrate observer feedback into RL

algorithms to learn legible (motion) trajectories. We add an observer that gives feedback to the agent to improve the legibility (see Section 5.2) of the learned policy. The interaction scheme of our proposed system is illustrated in Fig. 8.1.

This chapter relates to the bigger picture of this thesis by addressing our research question *How can we integrate actions that make use of social channel characteristics into RL?* (Q3). In our model the robot acts as teacher trying to communicate the goal as fast as possible opposed to only solving the task by executing pedagogical actions. The simulated observer acts as learner trying to infer the additional social signals provided by the pedagogical actions (see Fig. 6.7). The results presented in this chapter have been published in Bied and Chetouani (2020b).

8.2 Integrating Observer Feedback on Legibility into Interactive RL

In this work we are interested in the combination of a RL system with an observer that reasons about the goals of the learner to increase the legibility of the learned trajectories. In order to achieve this we use a MDP (see Section 3.4) in combination with reward shaping (see Section 3.4) to model the learning problem. We add the observer to the equation by modeling the observer with different strategies to estimate how likely the agent is going for the target goal.

8.2.1 Interactive RL

We formalize our problem by using a MDP defined as $(S, A, \mathcal{T}, R, \gamma)$ (see Section 3.4). A standard approach to solve problems formulated like this is Q-Learning. Q-Learning will also serve us as baseline to compare to. For Q-Learning we use a simple one-step Q-learning as defined by Eq. 3.6. For action taking we use the exploration rate ϵ, i.e. with a probability of ϵ the agent takes a random agent and the reward maximizing action of the current policy otherwise (see Section 3.4).

We add the observer to the system by using reward shaping Ng, Harada, and Russell (1999) (see Section 3.4). The original MDP reward is replaced by

$R'(s,a)$ by adding the weighted reward from the observer $\hat{O}$ to it.

$$R'(s,a) = R(s,a) + \beta \cdot \hat{O}(s,a) \tag{8.1}$$

The reward from the environment and the reward from the observer are of different nature and can, in general, differ in scale. The weighting factor β can be used to accommodate for this fact.

We will compare different algorithms for $\hat{O}$ to model different observer strategies. These algorithms will be presented in Section 8.2.3.

While other (more sophisticated) methods like policy shaping and value shaping (Knox and Stone, 2010; Knox and Stone, 2012b; Griffith et al., 2013; Najar, Sigaud, and Chetouani, 2019) (see Section 3.4) to integrate human feedback into the classical RL formulation exist, reward shaping will suffice as proof of concept for the feasibility of our approach.

Ng, Harada, and Russell (1999) describe the necessary requirements for reward shaping to preserve the optimal policy, if these requirements are not met positive-reward cycles can occur. Note that the way we are employing reward shaping does not meet these requirements.

8.2.2 Legibility

In this framework legibility serves as proxy how well the observer (learner) can understand the intentions of the agent (teacher) (see Fig. 6.3). In order to formally evaluate the legibility $\lambda(\xi)$ of a trajectory ξ, we use the legibility metric proposed by Dragan, Lee, and Srinivasa (2013) (see Section 5.2).

Following this line of work the observer needs to be able to confidently infer the correct (=the target) goal g^* after only observing a part of the whole trajectory to the goal $\xi_{s_0 \to s_t}$ starting at s_0 and ending at the intermediate point s_t. The trajectory is more legible the faster this confident inference happens.

Imagine an observer watching an agent acting in the environment shown in Fig. 8.1 in the observer model part. The observer tries now to infer as fast as possible for which goal the agent is going for. The right trajectory (solid blue line) is more legible than the left trajectory (dashed grey line), because for the right trajectory it seems more likely that the agent is going for the target goal

on the right. For the left trajectory it is still not clear for which goal the agent is going for, the next step could either be to the left or to the right.

Fig. 8.2 illustrates the concept of legibility in a discrete environment. The agent is aiming for the goal to the right (blue circle). There is an alternative goal on the left side (orange square). The more the trajectories go to the right side the higher is the resulting legibility. We will use this environment in our first experiment and refer to it as environment 1.

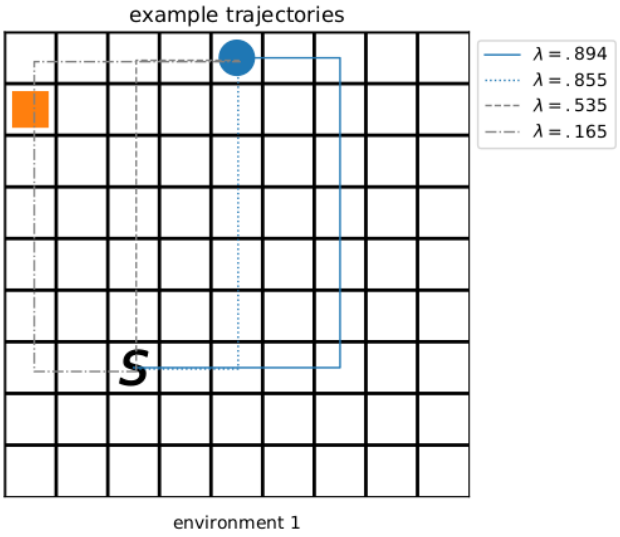


Figure 8.2: Different example trajectories with corresponding legibility (λ) in environment 1. The start position is marked with 'S', the target goal is marked with a blue circle and the alternative goal is marked with an orange square.

8.2.3 Modeling the Observer

We compare four algorithms: Q-Learning (Q-L), Q-OBS-D, Q-OBS-P and Q-OBS-L. These algorithms differ in the strategy the observer $\hat{O}$ implements. An overview of the ideas for the used functions for $\hat{O}$ is shown in Table 8.1.

The algorithms only differ in the choice of $\hat{O}$ inserted in Eq. 8.1. The main difference is between Q-Learning as baseline algorithm which we consider non-interactive and the other three algorithms which we consider interactive.

The main purpose of implementing different versions of the interactive methods is to explore how to integrate an observer that reasons about the possible

Table 8.1: The different used observer functions.

alg.	observer function idea
Q-L	Non-interactive baseline algorithm using no observer function
Q-OBS-D	Interactive algorithm using softmax of goal distance as observer function
Q-OBS-P	Interactive algorithm using the cost of the observed trajectory in comparison with the cost of the optimal trajectory as observer function
Q-OBS-L	Interactive algorithm using the legibility of the observed trajectory as observer function

goals of the agent in interactive RL. In the following we explain the proposed algorithms.

Q-L

Using the trivial equation for $\hat{O}$:

$$\hat{O} = 0 \tag{8.2}$$

is equivalent to plain Q-Learning. Q-Learning does not use any information from the observer and is therefore not interactive. Since Q-Learning only takes the rewards from the environment into account, it has no information on legibility.

However, this does not mean that the learned trajectories can not be legible, we can expect that some trajectories more legible than others. Therefore, Q-Learning will serve us as comparison to have a baseline how legible the trajectories are just by chance.

Q-OBS-D

$$\hat{O}(s, a, s') = \frac{1}{z} \exp(-\sigma d(s', g^*)) \tag{8.3}$$

d is the distance from s' to the goal using the Manhattan distance. z is partition function of the softmax distribution in order to normalize the probability to

one. σ is the temperature parameter to adjust how sharp the distribution peaks around the maximum.

Eq. 8.3 only depends on the current state s' and not on the observed trajectory snippet. We consider this approach as a naive approach to estimate goal probability and expect it to work in some cases, as it gives an incentive to reduce the distance to the target goal early on.

However, in more complex configurations, e.g. when the target goal is behind another goal, this approach might not work. Therefore, we expect it to work at least as good as Q-L, and in some cases even better.

Q-OBS-P

For Q-OBS-P we use the probability to reach a goal given a snippet of trajectory given with Eq. 5.9:

$$\hat{O}(\xi_{s_0 \to q}) = P(g^* | \xi_{s_0 \to q}) \tag{8.4}$$

Since this method uses a goal probability that has successfully been employed in previous research (Dragan, Lee, and Srinivasa, 2013; Dragan and Srinivasa, 2013; Holladay, Dragan, and Srinivasa, 2014) it seems like a more suitable candidate to estimate the goal probability than Q-OBS-D and we expect it to perform better.

Q-OBS-L

For Q-OBS-L we directly use the legibility as feedback from the observer. For the discrete case with K as the number of steps for reaching q and s_k as the state after k steps Eq. 5.2 becomes:

$$\hat{O}(\xi_{s_0 \to q}) = \frac{\Sigma_{k=0}^{K} P(g | \xi_{s_0 \to s_k}) f(k)}{\Sigma_{k}^{K} f(k)} \tag{8.5}$$

Using directly the legibility is not a goal probability, since it does not sum up to one for all goals, nevertheless it contains by definition information on how confident the observer is that the agent is going for the target goal. Therefore, we also expect this method to also perform better than Q-OBS-D.

8.3 Experiments

The goal of the experiments is to evaluate the ability of the algorithms presented in Section 8.2.3 to increase the legibility of the learned trajectories.

Q-Learning will serve as non-interactive baseline to compare to. Q-OBS-D, Q-OBS-P and Q-OBS-L integrate information information on the goal probability into the model and are expected to perform better.

We evaluated the approach on five different environments. For the first environment there are only two possible goals, and we use it to illustrate the approach. For the environments 2 – 5 we use three goals and changed the configuration of these goals relative to each other.

The parameters were set intuitively. First we set the parameters that Q-Learning performed reasonable well and kept these parameters for the interactive algorithms. The parameters specific to the interactive algorithms were then set to perform reasonable well, but not tweaked to achieve the best possible performance. The parameters were kept for all environments.

For the rewards from the environment we used: reaching the target goal $r_g = 0$, penalty for unvisited state different from the target goal $r_p = -0.1$, penalty for already visited state different from the target goal $r_{p2} = -0.2$.

For the Q-Learning relevant parameters we used: $\alpha = 0.9$, $\gamma = 0.9$ and $\epsilon = 0.1$. The q-table was initialized with random values from 0 to 2.

For Q-OBS-D we set $\sigma = 0.3$. For implementation reasons, to address the problem of positive loops we use $\beta = \beta_1\beta_2$, with $\beta_1 = -r_p$ and $\beta_2 = 2$. By setting the parameters like this, we assure that agent does not achieve a net gain larger than 0 by cycling back and forth. However, the possible looping behavior drastically limits the choice to set β.

The used parameters are compactly shown in Table 8.2. Each algorithm was trained in 100 sessions for 120 episodes on each environment.

parameter	value
r_g	0
r_p	-0.1
r_{p2}	-0.2
α	0.9
γ	0.9
ϵ	0.1
β	$\beta_1\beta_2$
β_1	$-r_p$
β_2	2

Table 8.2: Parameter and corresponding values used in the experiments.

8.3.1 Environment 1

Description

The first environment (see Fig. 8.2) was used to check the feasibility of the approach and includes only two goals: the target goal and one alternative goal. The size of the grid of the first environment is 9x9 and is visualized in Fig. 8.2 alongside with four example trajectories and the corresponding legibility.

The first trajectory (from left to right) is sup-optimal in terms of steps towards the target goal and the legibility is low, the second and the third trajectory are both optimal, however the third trajectory yields a higher legibility because one can infer earlier for which goal the agent is aiming. The fourth trajectory is sup-optimal but the legibility is the highest of the shown trajectories.

There are multiple optimal trajectories, when using Q-Learning, there is no reason for the agent to prefer one optimal trajectory over another optimal trajectory. Since the learning is stochastic, we expect the agent to sometimes learn an optimal trajectory with a higher legibility and other times with a lower legibility.

We do not expect to learn with Q-Learning trajectories with a even higher legibility. When integrating the observer feedback, we expect the learned trajectories to be more legible and sometimes to even learn trajectories that are sub-optimal, but more legible than the most legible optimal trajectory. While we show only two possible optimal trajectories to the goal, there are more possible optimal trajectories to the goal than these two. These trajectories

only differ in the legibility. Since the Q-table is randomly initialized and an ϵ-greedy exploration strategy is used the learning process is stochastic.

Note that technically we are not learning trajectories, but policies - once a policy is learned the trajectory generated by that policy are deterministic. When speaking about the learned trajectories, we are strictly speaking about the trajectories that are generated by the learned policies.

Results

As aforementioned, even when only considering only the optimal trajectories there is a large number of possible trajectories. During the training processes of the different algorithms a large number of different trajectories have been learned. It is not possible to visualize the differences of the different algorithms in only one graph. Therefore we will use different methods to illustrate the occurred differences.

First, we will have a look into the five best and five worst trajectories w.r.t. the legibility as illustrated in Fig. 8.3.

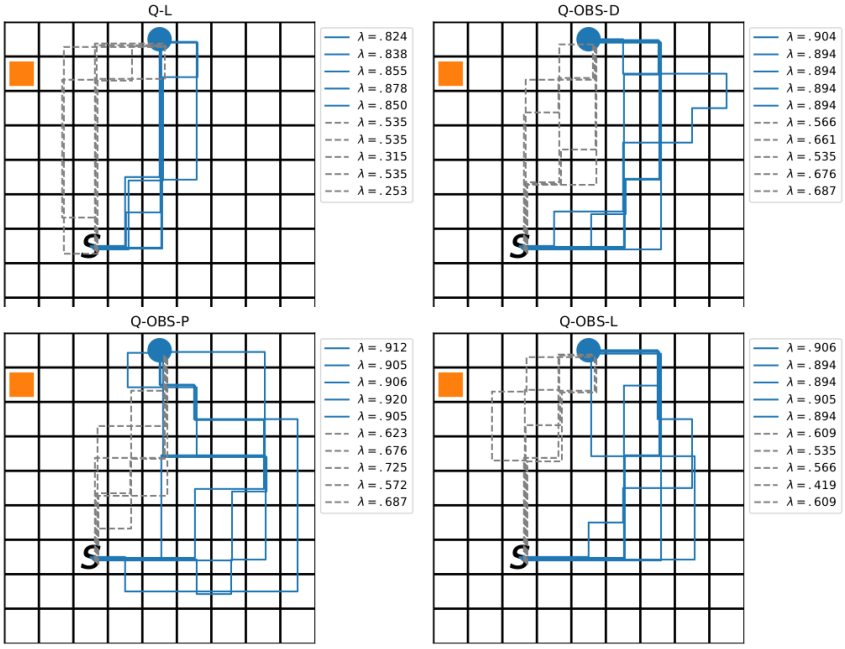


Figure 8.3: The five best (solid blue lines) and five worst (dashed grey lines) learned trajectories w.r.t. λ for the different algorithms of environment 1.

If we now have a look at the legibility of the best and worst trajectories (w.r.t.

λ) learned by Q-Learning, we see that these values are lower than legibility of the best and worst trajectories learned by the interactive algorithms.

From Fig. 8.2 we know that the more legible trajectories tend to go to the right earlier on and are more on the right side of the grid in general. We can see that the trajectories of the interactive algorithms also tend to lie more on the right side of the grid world.

Next, we will have a look onto the heatmap of the learned trajectories in Fig. 8.4. We can see that not only the best and worst trajectories for the interactive algorithms lie more to the right, but also in the heatmap the interactive algorithms are 'hotter' in the regions of the more legible trajectories.

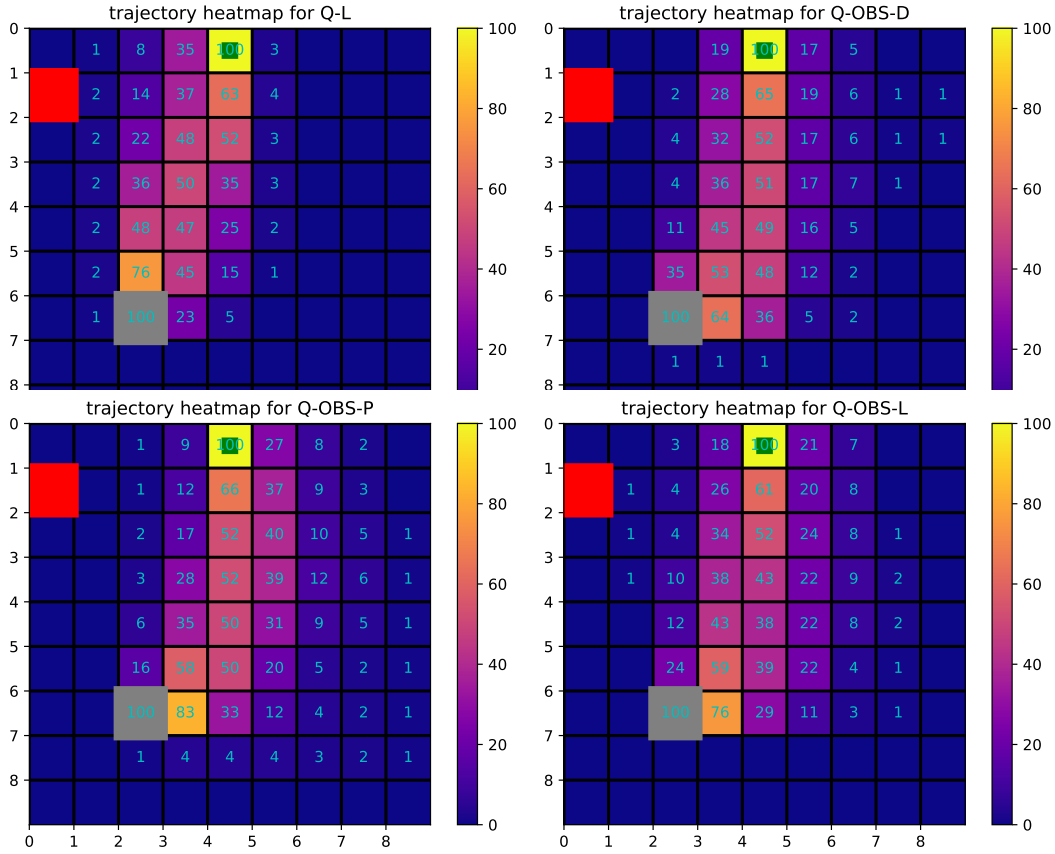


Figure 8.4: Heat map of the learned trajectories for the different algorithms in environment 1.

The legibility of the different algorithms averaged over 100 runs for environment 1 is reported in Fig. 8.5. All algorithms that integrate a non-zero observer reward perform significantly better than plain Q-Learning. Q-OBS-P performs

best regarding the legibility of the learned trajectories.

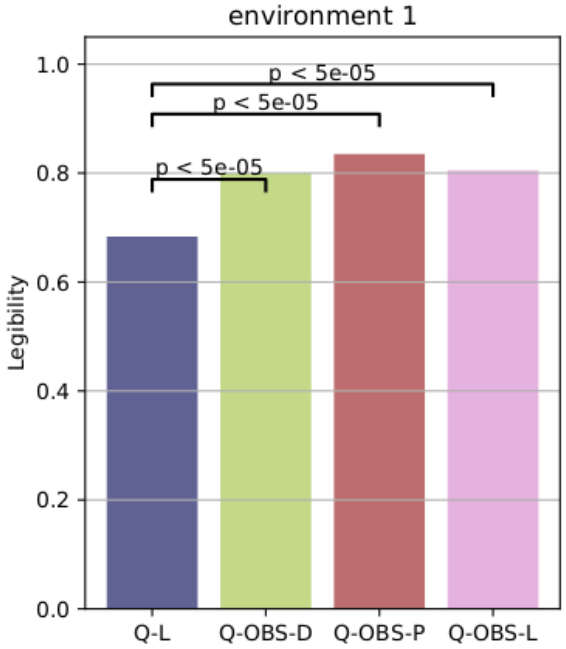


Figure 8.5: Mean of the legibility for the different algorithms in environment 1. The significance level was calculated using the Mann-Whitney U test.

8.3.2 Environments 2 – 5

While we showed in environment 1 that all interactive algorithms perform better than Q-Learning, we tested the approach on four additional environments to test the limits of our approach. This time we included an additional alternative goal. The grid size for tasks 2 – 5 is 9x9 as in environment 1.

All tasks have three goals, the target goal and two alternative goals. The different environments can be seen in Fig. 8.6. We varied the relative configuration of the goals to evaluate the influence on the performance of the algorithms.

In environment 2 – 4 the position of the alternative goals stays the same, we only vary the position of the target goal. In environment 5 there is no obvious more legible trajectory, so we do not expect including the observer feedback to perform better.

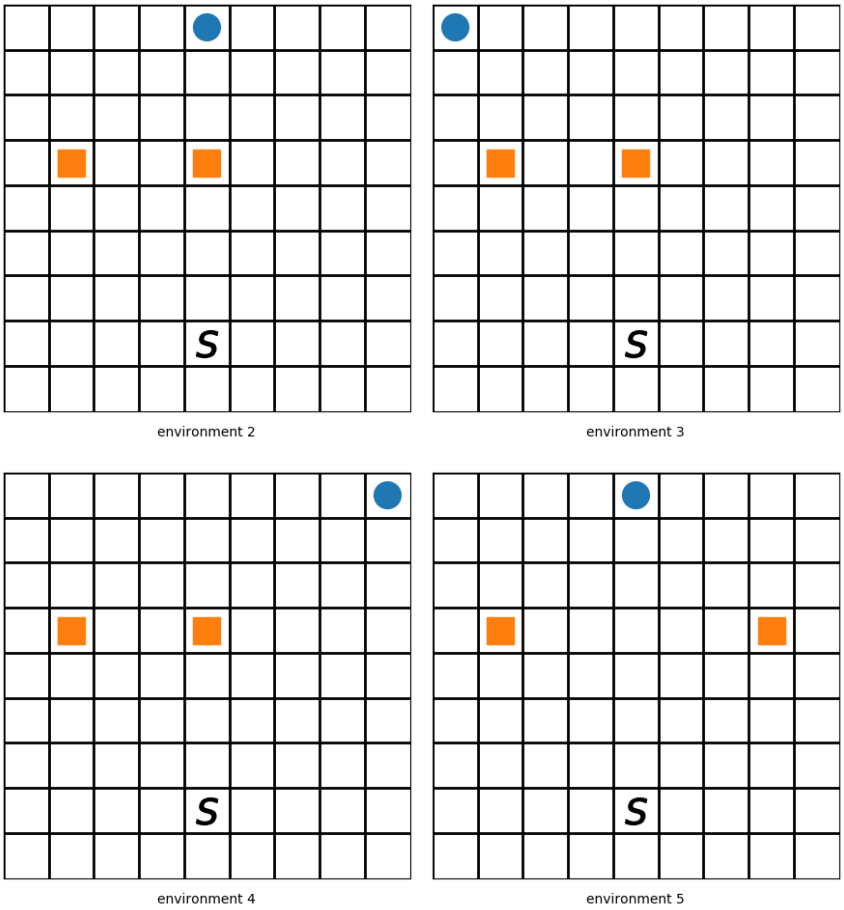


Figure 8.6: Environment 2 – 5, for the environments 2 – 4 only the position of the target goal was varied, for environment 5 there is no legible path from a human point of view.The target goal is marked as a blue circle, the alternative goals as orange squares and the start with 'S'.

Results

The legibility of the different algorithms averaged over 100 runs for environment 2 – 5 is reported in Fig. 8.7. While Q-OBS-D significantly improved the legibility of the learned trajectory for environment 1, there is no significant difference for environment 2 and 4.

As for environment 1, Q-OBS-P is the best performing algorithm for all environments with a significant higher legibility in comparison to Q-Learning. Q-OBS-P performs significantly better than Q-Learning for environments 2 – 4, but not for environment 5.

For environment 5, from a human perspective, there is no more legible trajectory

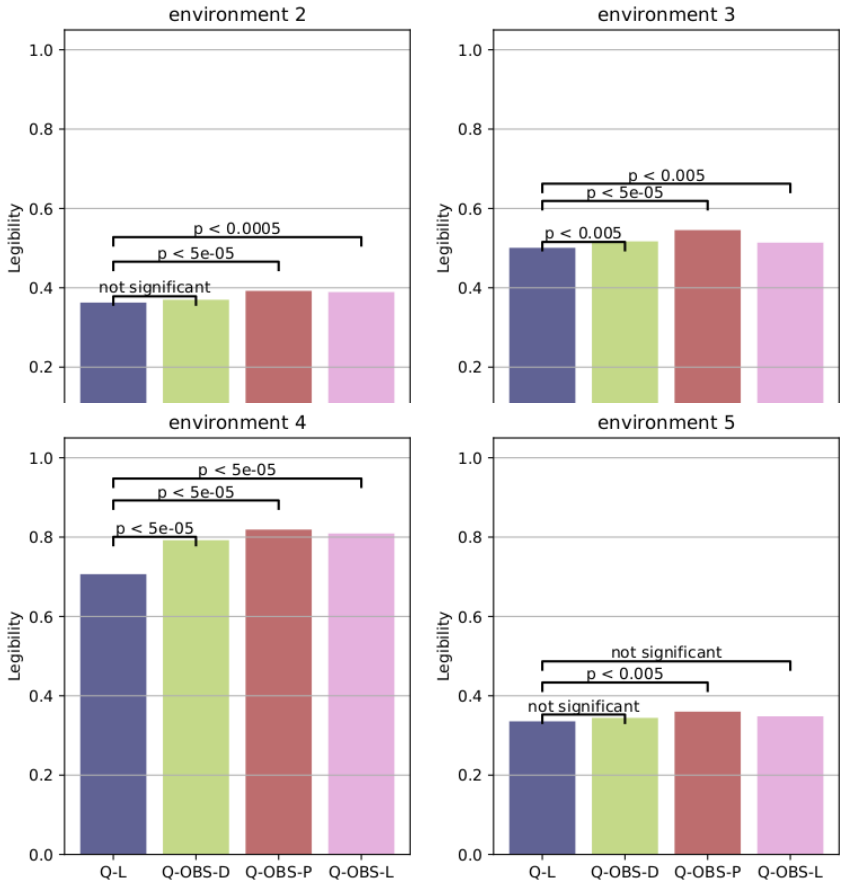


Figure 8.7: Mean of the legibilities for Task 2 – 5 for the different algorithms. The significance level was calculated using the Mann-Whitney U test.

than the (only) optimal trajectory i.e. just going straight from start to the target goal. The only optimal trajectory has a legibility of $\lambda = 0.388$. In Fig. 8.8 we see the five best and five worst trajectories for environment 5. We see that in terms of the metric that all algorithms generated some trajectories with $\lambda > 0.388$.

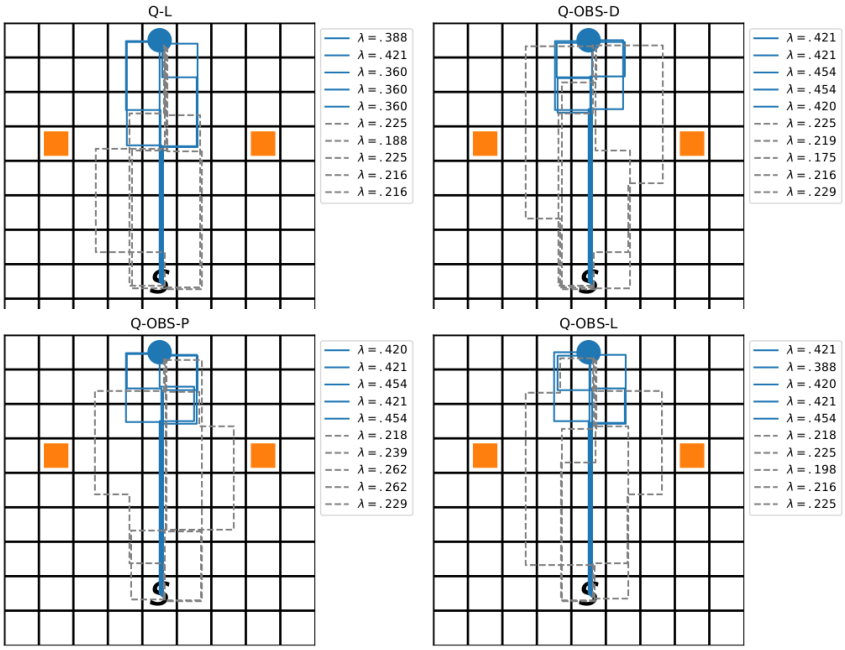


Figure 8.8: The five best (solid blue lines) and five worst (dashed grey lines) learned trajectories w.r.t. λ for the different algorithms of environment 5.

8.4 Discussion

The results show that the interactive algorithms perform better than Q-Learning. Our main focus is on showing that the interactive approaches are useful in comparison to the non-interactive approaches. In order to support our main message it is not really important which of the interactive algorithm performs best.

A major limitation is the use of reward shaping and a next step will be to replace it with a better suited method like policy shaping. Therefore it is not useful to put effort into analyzing differences in the approaches based on reward shaping, especially since the parameters were not tuned for every algorithm to perform to its best.

In our approach we simulated the observer giving additional rewards to the agent. One possible idea is to employ a real human in the loop giving the observer feedback. Research on human feedback in RL (e.g. Thomaz and Breazeal, 2006a; Thomaz and Breazeal, 2008; Ho, Littman, Cushman, et al., 2015; Ho, Cushman, et al., 2019) suggests that probably real humans will behave differently than our models, therefore the framework might not

work.

However, our model might be useful even when no observer giving feedback is present. Since the agent has all the information the observer has, we could integrate the observer model internally into the agent. The agent could improve it's behavior by expecting to being watched. An approach like this would emphasize the idea of ToM for the agent.

One downside of our approach is that the observe needs to know all the present goals in the setting to infer the goal probability. In robotics this is a strong assumption. An interesting problematic arises in this context: the robustness of the legibility if the observer has only partial knowledge of the goals.

Also interesting is that in environment 5 there are more legible trajectories than the trajectory that goes directly from start to the target goal. From a human point of view there are arguably no more legible trajectories. For the legibility metric this happens, because for example going to the right after passing the level of the two alternative goals, drastically decreases the probability of the left goal and increases the probability of the right goal, therefore the target goal probability also increases. Simultaneously the length of the trajectory increases leading to a change of the weights for each part of the distribution. These two changes together can lead to an increase in legibility.

It is not clear, if this will be a relevant problem, for longer continuous trajectories. That's another limitation we did not address in this work - scaling the approach up to a more complex task than just a grid world, possibly also using a robot with multiple degrees of freedom instead of just a point robot.

8.5 Conclusion

In this work, we were interested in integrating observer feedback into RL to increase the legibility of the learned trajectories. We proposed three interactive RL algorithms by integrating observer feedback and compared them to the non-interactive Q-Learning. We showed that the interactive RL approaches learn trajectories with a significantly higher legibility and that even a simple approach can perform at least as good as Q-Learning.

From that, we conclude that when it comes to Human-Robot cooperation it is useful to integrate reasoning about the goal probabilities in order to increase

the legibility of the trajectories. While we used reward shaping as a simple mechanism to integrate the feedback, the problem of positive-reward cycle is limiting the power of the approach.

Furthermore, the research in this chapter serves as answer to our research question *How can we integrate actions that make use of social channel characteristics into RL?* (Q3). By learning to produce more legible trajectories, the agent augments its actions with a part that communicates the goal as social signal. Thus, we have shown, how to integrate actions that use social channel characteristics into RL.

The work presented in this chapter could be extended by considering other shaping mechanisms, as for example policy shaping, as it will probably work better in experiments with real humans. Possible directions include more complex environments and experiments with humans.

Chapter

9 Discussion and Conclusion

Contents

9.1 Summary of Contributions

In this thesis we are interested how signals that combine task and social signals in one signal can be used in HRI settings. We focus on the sensorimotor channel, since it is meaningful for robotics and allows to combine task and social signals.

More specific, we are interested in pedagogical situations, where a teacher has the intention to communicate additional information to a learner. We are interested in the human side, as well in the robot side of the HRI.

We are interested in how humans are using the sensorimotor channel when solving a task in contrast to teaching the task to a robot, but also how sensorimotor actions are perceived by humans. Further, we are interested in how these kind of actions can be implemented on the robot side.

In order to give more structure to research focusing on channel usage in the

context of HRI and put our own research into perspective we propose a general communication model for HRI that explicitly distinguishes between task-, social- and combined signals, as well as specific model describing the combined channel usage in pedagogical situations in more detail. The first contribution are the proposed models for communication in HRI. While we do not claim that the ideas implemented in these models are novel, combining them in an explicit models contributes by providing a base to guide future research on the channel usage in HRI.

The second contribution are the insights on human behavior and perception of sensorimotor actions in the context of HRI. Namely, humans use the possibility to include additional information in their actions when teaching a robot how to solve a task in contrast to just solve it. The demonstrations coming from a teaching setting are perceived as more informative than when the task is just solved. Additionally, humans perceive the use of negative demonstrations as more informative than positive demonstrations. Further, when humans are only given the possibility to teach the task via only one demonstration they tend to just solve the task, as opposed to when they can give multiple demonstrations where they tend to include additional information in the demonstrations.

The third contribution is the proposition of a framework based on RL that integrates reasoning about a potential observer into the learning process. The framework allows to learn trajectories that are more legible for an observer. Here a higher legibility serves a an approximation for a higher level of social signals apart from the task signals provided by the actions. Further, we compare different algorithms to implement the observer reasoning and show that all algorithms that include reasoning about the observer achieve higher legibility than classical RL baseline.

9.2 General Limitations of the Approach

A general limitation of this approach is that by focusing only on one channel we do not now the influence of adding other channels. Adding other communication channel might strongly impact how humans use the sensorimotor channel. For example in work using the Sophie's kitchen framework (Thomaz and Breazeal, 2008; Thomaz and Breazeal, 2006b), humans used the feedback channel to also give motivational feedback. This effect vanished as soon as an

explicit channel to give motivational feedback was added.

Thus, already changing even small things in the interaction protocol might have strong influences on the human channel usage. Adding additional channels might completely change the usage of the already present channels in a protocol. This change in channel usage can be a disadvantage as well as an advantage. A disadvantage is that research results found while investigating only a single channel isolation might be invalidated when adding another channel. An advantage would be when an overlay of how humans use a channel is detected that is difficult to untangle. In this case the introduction of a new channel explicit for the function how humans overlaid the original channel can solve the problem.

Another problem that comes with the inference of intentions is that the inference might only be partially correct or fail completely. Humans have the capability (at least so some extend) to recognize if their intentions were correctly understood. If not they have plenty of mechanisms recover when their intentions are not understood. They have the capability of changing the interaction protocol during an ongoing interaction and modify as they see it fit. If the new behavior is not understood again and the interaction partner reacts in way they have not foreseen they can adapt again.

However, the problem that applies to HRI in general, and to our approach in particular, is that the protocol of the interaction loop, has to be well defined prior the interaction. The protocol can not 'completely change' during an interaction. These protocols can integrate reaction options for foreseen communication problems, but they can not completely change.

9.3 Perspectives

The reflections that revolve around the various contributions of our research work and the resulting perspectives can be summarized as follows:

- **Goal inference instead of obstacle inference**: In our user study, we designed the task in a way that the goal was known and an obstacle in the way should be learned. In our RL framework, we changed to goal inference instead of obstacle inference, because legibility provided us a theoretical metric that has already been tested in HRI. In hindsight it

would make sense to conduct the user study where the learning task consists of inferring the goal instead of the obstacle. This would not only increase consistency, but likely generate additional interesting insights.

- **Learning from human data:** As a first step of the LfD pipeline we collected human demonstrations in our user study. However, we did not implement the subsequent steps and did not use the data to learn a generative model. Since we did an exploratory study to get insights how different aspects (number of demonstrations, positive/negative demonstrations, different task instances) influence the behavior in combination that human behavior differs from person to person even if all variables are kept the same, we could not collect enough data to learn from it. While approaches that focus on one shot learning (e.g. DMP), we would need to integrate the different variable aspects as prior knowledge. Integrating this prior knowledge of different kind is not a simple task.
- **Automatic classification:** Similar to the before mentioned reason, while we collected enough data to show a statistical significant difference between examples from the solving and teaching condition, we could not train a classifier to automatically distinguish between the conditions. An interesting direction for a data set collected in a less exploratory manner would be to train classifiers to automatically detect data from the teaching condition.
- **User study with a real robot:** In our user study, the task was to be executed on a tablet. The user did not directly interact with a real robot, but were introduced to a pepper robot with a picture and told that this robot should later learn from their demonstrations. Human teaching behavior will most likely be influenced by how capable humans consider the robot. Thus, interacting with a real robot might have an influence on the results. However, executing a study where participants with little to no knowledge about robots teach physical tasks to a real robot increases the complexity of the experiment by a lot.
- **Negative demonstrations:** In our work we did not continue in the direction of implementing learning from negative demonstrations into LfD, since we worked on RL instead. However, this is a promising direction, as we showed that negative demonstrations are perceived as informative and

useful by humans. While for example Cakmak and Thomaz (2011) and Calinon (2019) mention negative demonstrations as a good direction, this approach has not really been addressed in LfD research.

- **Learning from humans with pedagogical intentions:** We showed that humans include additional information in their demonstrations when teaching a robot. Thus, it seems reasonable that demonstrations from teaching conditions can be helpful to support the learning of a robot. However, this might not be straight forward: as Milli and Dragan (2019) show, even when humans try to be pedagogical in terms of learning it's safer to assume that they give literal demonstrations.
- **Employ the RL framework with a more complex task and a real robot:** We applied our approach to a small discrete task and used a point-robot. An extension of the work could investigate how the approach to integrate reasoning about a potential observer can be applied with a real robot for a task that is closer to a real life application. However, for this the value-based approach does not seem appropriate, since value-based approaches tend to do not deal well with high-dimensionality.
- **User study employing the RL framework:** In our RL framework we did not interact with a real human, but only with a simulated observer. To measure success we used the legibility metric as proxy how well an observer can infer the goals of the actions. Further directions include either replacing the simulated observer with human participants or testing how fast human participants can infer the agents goals, respectively a combination of both. However, the first approach seems unfeasible, as the number of interactions will be too high.
- **Advancing to more sophisticated shaping methods:** In order to integrate the observer feedback into RL we used reward shaping a simple method. Reward shaping can lead to infinite loops, while we did not encounter infinite loops, it restricted the choice of our parameters. Implementing a more sophisticated method like policy shaping could solve this problem. Moreover, in an interaction with humans, policy shaping might be better suited in interactions with humans.

9.4 Conclusion

In this thesis we conducted research on actions that provide combined task and social signals in one signal in the context of HRI. We showed that these kind of actions are used by humans when teaching a robot how to solve a task, and that humans perceive these actions as informative in HRI.

Further, we showed that by integrating reasoning about the interaction partner we can integrate actions that provide combined signals into RL to learn trajectories that are more legible to an observer.

We applied our research only to a small set of task, and thus could only contribute with one step towards robots that can emulate human-like capabilities. However, the results of our research indicate that advancing research into a direction of exploiting combined actions, in particular sensorimotor actions, is a promising direction to go for HRI.

Nevertheless, in order to fulfill a long time human vision and release robots into the wild and have them interact with humans and their environment will require a lot more research to be done that can only be done step by step.

Bibliography

Abbeel, Pieter, Adam Coates, and Andrew Y. Ng (2010). "Autonomous Helicopter Aerobatics through Apprenticeship Learning". In: *The International Journal of Robotics Research* 29.13, pp. 1608–1639. DOI: 10.1177/0278364910371999.

Akgun, Baris, Maya Cakmak, Jae Wook Yoo, and Andrea L. Thomaz (2012). "Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective". In: *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 391–398. DOI: 10.1145/2157689.2157815.

Argall, Brenna D., Sonia Chernova, Manuela Veloso, and Brett Browning (2009). "A survey of robot learning from demonstration". In: *Robotics and Autonomous Systems* 57.5, pp. 469–483. ISSN: 09218890. DOI: 10.1016/j.robot.2008.10.024.

Asimov, Isaac (1988). *The complete robot*. Garden City, N.Y.: Doubleday.

Atkeson, C.G. and Stefan Schaal (1997). "Learning tasks from a single demonstration". In: *Proceedings of International Conference on Robotics and Automation*. Vol. 2, 1706–1712 vol.2. DOI: 10.1109/ROBOT.1997.614389.

Bagnell, J. Andrew and J.G. Schneider (2001). "Autonomous helicopter control using reinforcement learning policy search methods". In: *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164)*. Vol. 2, 1615–1620 vol.2. DOI: 10.1109/ROBOT.2001.932842.

Baker, Chris L., Rebecca Saxe, and Joshua B. Tenenbaum (2009). "Action understanding as inverse planning". In: *Cognition* 113, pp. 329–349.

Bandura, Albert and David C. McClelland (1977). *Social learning theory*. Vol. 1. Englewoodcliffs Prentice Hall.

Barto, Andrew G., Richard S. Sutton, and Charles W. Anderson (1983). "Neuronlike adaptive elements that can solve difficult learning control problems". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13.5, pp. 834–846. DOI: 10.1109/TSMC.1983.6313077.

Belpaeme, Tony, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka (Aug. 2018). "Social robots for education: A review". In: *Science Robotics* 3.21. DOI: 10.1126/scirobotics.aat5954.

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston (2009). "Curriculum Learning". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. Montreal, Quebec, Canada: ACM, pp. 41–48. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553380.

Bied, Manuel and Mohamed Chetouani (2020a). "Exploring the Difference between Solving and Teaching in Sensorimotor Tasks". In: *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '20. Cambridge, United Kingdom: Association for Computing Machinery, pp. 139–141. ISBN: 9781450370578. DOI: 10.1145/3371382.3378284.

– (2020b). "Integrating an Observer in Interactive Reinforcement Learning to Learn Legible Trajectories". In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 760–767. DOI: 10.1109/RO-MAN47096.2020.9223338.

Billard, Aude, Sylvain Calinon, Ruediger Dillmann, and Stefan Schaal (2008). *Survey: Robot Programming by Demonstration*. Handbook of Robotics, Chapter 59.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387310738.

Bizzi, E., A. D'Avella, P. Saltiel, and M. Tresch (2002). "Book Review: Modular Organization of Spinal Motor Systems". In: *The Neuroscientist* 8.5. PMID: 12374428, pp. 437–442. DOI: 10.1177/107385802236969.

Bobu, Andreea, Marius Wiggert, Claire J. Tomlin, and Anca D. Dragan (2020). "Feature Expansive Reward Learning: Rethinking Human Input". In: *CoRR* abs/2006.13208.

Breazeal, Cynthia, Matt Berlin, Andrew Brooks, Jesse Gray, and Andrea L. Thomaz (2006). "Using perspective taking to learn from ambiguous demonstrations". In: *Robotics and Autonomous Systems* 54.5. The Social Mechanisms of Robot Programming from Demonstration, pp. 385–393. ISSN: 0921-8890. DOI: https://doi.org/10.1016/j.robot.2006.02.004.

Breazeal, Cynthia and Andrea L. Thomaz (2008). "Learning from human teachers with Socially Guided Exploration". In: *2008 IEEE International Conference on Robotics and Automation*, pp. 3539–3544. DOI: 10.1109/ROBOT.2008.4543752.

Broekens, Joost and Mohamed Chetouani (2019). "Towards Transparent Robot Learning through TDRL-based Emotional Expressions". In: *IEEE Transactions on Affective Computing*, pp. 1–1. ISSN: 2371-9850. DOI: 10.1109/TAFFC.2019.2893348.

Brooks, Rodney, Cynthia Breazeal, Matthew Marjanovic, Brian Scassellati, and Matthew Williamson (Mar. 2002). "The Cog Project: Building a Humanoid Robot". In: *Lecture Notes in Artificial Intelligence* 1562. DOI: 10.1007/3-540-48834-0_5.

Burke, Christopher J., Philippe N. Tobler, Michelle Baddeley, and Wolfram Schultz (2010). "Neural mechanisms of observational learning". In: *Proceedings of the National Academy of Sciences* 107.32, pp. 14431–14436. ISSN: 0027-8424. DOI: 10.1073/pnas.1003111107.

Bütepage, Judith and Danica Kragic (2017). "Human-Robot Collaboration: From Psychology to Social Robotics". In: *CoRR* abs/1705.10146.

Cakmak, Maya and Manuel Lopes (2012a). "Algorithmic and Human Teaching of Sequential Decision Tasks". In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI'12. Toronto, Ontario, Canada: AAAI Press, pp. 1536–1542.

– (2012b). "Algorithmic and human teaching of sequential decision tasks". In: *AAAI 2012*.

Cakmak, Maya and Andrea L. Thomaz (2011). "Active Learning with Mixed Query Types in Learning from Demonstration". In: *ICML Workshop on New Developments in Imitation Learning*.

– (2014). "Eliciting good teaching from humans for machine learners". In: *Artificial Intelligence* 217, pp. 198–215. ISSN: 00043702. DOI: 10.1016/j.artint.2014.08.005.

Calinon, Sylvain (Sept. 2015a). "A Tutorial on Task-Parameterized Movement Learning and Retrieval". In: *Intelligent Service Robotics* 9. DOI: 10.1007/s11370-015-0187-9.

– (2015b). "Robot Learning with Task-Parameterized Generative Models". In: *ISRR*.

– (2019). "Learning from Demonstration (Programming by Demonstration)". In: *Encyclopedia of Robotics*. Ed. by M. H. Ang, O. Khatib, and B. Siciliano. Springer. DOI: 10.1007/978-3-642-41610-1_27-1.

Calinon, Sylvain and Aude Billard (2007). "Active teaching in robot programming by demonstration". In: *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pp. 702–707. DOI: 10.1109/ROMAN.2007.4415177.

Calinon, Sylvain, Florent Guenter, and Aude Billard (2007). "On learning, representing, and generalizing a task in a humanoid robot". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*. ISSN: 10834419. DOI: 10.1109/TSMCB.2006.886952.

Carter, Elizabeth J., Jessica K. Hodgins, and David H. Rakison (2011). "Exploring the neural correlates of goal-directed action and intention understanding". In: *NeuroImage* 54.2, pp. 1634–1642. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2010.08.077.

Castellano, Ginevra, Iolanda Leite, and Ana Paiva (June 2017). "Detecting Perceived Quality of Interaction with a Robot Using Contextual Features". In: *Auton. Robots* 41.5, pp. 1245–1261. ISSN: 0929-5593. DOI: 10.1007/s10514-016-9592-y.

Castellano, Ginevra, André Pereira, Iolanda Leite, Ana Paiva, and Peter W. McOwan (2009). "Detecting User Engagement with a Robot Companion Using Task and Social Interaction-Based Features". In: *Proceedings of the 2009 International Conference on Multimodal Interfaces*. ICMI-MLMI '09. Cambridge, Massachusetts, USA: Association for Computing Machinery, pp. 119–126. ISBN: 9781605587721. DOI: 10.1145/1647314.1647336.

Celemin, Carlos and Javier Ruiz-del-Solar (July 2015). "COACH: Learning continuous actions from COrrective Advice Communicated by Humans". In: *2015 International Conference on Advanced Robotics (ICAR)*, pp. 581–586. DOI: 10.1109/ICAR.2015.7251514.

Celemin, Carlos, Javier Ruiz-del-Solar, and Jens Kober (2019). "A fast hybrid reinforcement learning framework with human corrective feedback". In: *Autonomous Robots* 43, pp. 1173–1186.

Chakraborti, Tathagata, Anagha Kulkarni, Sarath Sreedharan, David E. Smith, and Subbarao Kambhampati (2018). "Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior". In: *CoRR* abs/1811.09722.

Chaminade, Thierry and Gordon Cheng (2009). "Social cognitive neuroscience and humanoid robotics". In: *Journal of Physiology-Paris* 103.3. Neurorobotics, pp. 286–295. issn: 0928-4257. doi: https://doi.org/10.1016/j.jphysparis.2009.08.011.

Chaveroche, Maxime, Adrien Malaisé, Francis Colas, François Charpillet, and Serena Ivaldi (2018). *A Variational Time Series Feature Extractor for Action Prediction.*

Chen, Nutan, Justin Bayer, Sebastian Urban, and Patrick van der Smagt (Nov. 2015). "Efficient movement representation by embedding Dynamic Movement Primitives in Deep Autoencoders". In: doi: 10.1109/HUMANOIDS.2015.7363570.

Chen, Nutan, M. Karl, and P. van der Smagt (2016). "Dynamic movement primitives in latent space of time-dependent variational autoencoders". In: *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pp. 629–636.

Chernova, Sonia and Andrea L. Thomaz (2014). *Robot Learning from Human Teachers*. Morgan & Claypool Publishers. isbn: 1627051996.

Cruz, Francisco, Johannes Twiefel, Sven Magg, Cornelius Weber, and Stefan Wermter (2015). "Interactive reinforcement learning through speech guidance in a domestic scenario". In: *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. doi: 10.1109/IJCNN.2015.7280477.

Csibra, Gergely and György Gergely (2006). "Social Learning and Social Cognition: The Case for Pedagogy". In: *Progress of Change in Brain and Cognitive Development. Attention and Performance XXI*, pp. 249–274.

– (2007). "'Obsessed with goals': functions and mechanisms of teleological interpretation of actions in humans." In: *Acta psychologica* 124 1, pp. 60–78.

– (2009). "Natural pedagogy". In: *Trends in Cognitive Sciences* 13.4, pp. 148–153. issn: 1364-6613. doi: https://doi.org/10.1016/j.tics.2009.01.005.

Cui, Yuchen and Scott Niekum (2018). "Active Reward Learning from Critiques". In: *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 6907–6914. ISBN: 9781538630815. DOI: 10.1109/ICRA.2018.8460854.

Dautenhahn, Kerstin (2007). "Methodology & themes of human-robot interaction: A growing research field". In: *International Journal of Advanced Robotic Systems*. ISSN: 17298806.

Dautenhahn, Kerstin and Chrystopher L. Nehaniv (2002). *Imitation in Animals and Artifacts*. Cambridge, MA, USA: MIT Press. ISBN: 0-262-04203-7.

Deisenroth, Marc, Gerhard Neumann, and Jan Peters (2013). "A Survey on Policy Search for Robotics". In: *Foundations and Trends® in Robotics* 2.1–2, pp. 1–142. ISSN: 1935-8253. DOI: 10.1561/2300000021.

Deisenroth, Marc and Carl Rasmussen (Jan. 2011). "PILCO: A Model-Based and Data-Efficient Approach to Policy Search." In: pp. 465–472.

Dennett, Daniel C. (1987). *The Intentional Stance*. MIT Press.

Dockendorff, Martin, Natalie Sebanz, and Guenther Knoblich (2019). "Deviations from optimality should be an integral part of a working definition of SMC: Comment on "The body talks: Sensorimotor communication and its brain and kinematic signatures" by Pezzulo et al." In: *Physics of life reviews* 28, pp. 22–23.

Dragan, Anca D., Kenton C. T. Lee, and Siddhartha S. Srinivasa (Mar. 2013). "Legibility and predictability of robot motion". In: *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 301–308. DOI: 10.1109/HRI.2013.6483603.

Dragan, Anca D. and Siddhartha S. Srinivasa (July 2012). "Formalizing Assistive Teleoperation". In: *Proceedings of Robotics: Science and Systems*.

– (2013). "Generating Legible Motion". In: *Robotics: Science and Systems*.

Fisac, Jaime F., Chang Liu, Jessica B. Hamrick, S. Shankar Sastry, J. Karl Hedrick, Thomas L. Griffiths, and Anca D. Dragan (2018). "Generating Plans that Predict Themselves". In: *CoRR* abs/1802.05250.

Flash, Tamar and Binyamin Hochner (2005). "Motor primitives in vertebrates and invertebrates". In: *Current Opinion in Neurobiology* 15.6. Motor sytems / Neurobiology of behaviour, pp. 660–666. ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2005.10.011.

Fong, Terrence, Illah Nourbakhsh, and Kerstin Dautenhahn (Mar. 2003). "A Survey of Socially Interactive Robots". In: *Robotics and Autonomous Systems* 42, pp. 143–166. doi: 10.1016/S0921-8890(02)00372-X.

Fournier, Pierre, Olivier Sigaud, and Mohamed Chetouani (Aug. 2017). "Combining artificial curiosity and tutor guidance for environment exploration". In: *Workshop on Behavior Adaptation, Interaction and Learning for Assistive Robotics at IEEE RO-MAN 2017*. Special Issue based on the 2nd Workshop on Behavior Adaptation, Interaction and Learning for Assistive Robotics at IEEE RO-MAN 2017. Lisbon, Portugal.

Gergely, György, Zoltán Nádasdy, Gergely Csibra, and Szilvia Bíró (1995). "Taking the intentional stance at 12 months of age". In: *Cognition* 56.2, pp. 165–193. issn: 0010-0277. doi: https://doi.org/10.1016/0010-0277(95)00661-H.

Goldman, S.A. and M.J. Kearns (Feb. 1995). "On the Complexity of Teaching". In: *J. Comput. Syst. Sci.* 50.1, pp. 20–31. issn: 0022-0000. doi: 10.1006/jcss.1995.1003.

Grice, H. Paul (1957). "Meaning". In: *Philosophical Review* 66.3, pp. 377–388. doi: 10.2307/2182440.

Griffith, Shane, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L. Thomaz (2013). "Policy Shaping: Integrating Human Feedback with Reinforcement Learning". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Curran Associates, Inc., pp. 2625–2633.

Grondman, Ivo, Lucian Busoniu, Gabriel A. D. Lopes, and Robert Babuska (2012). "A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.6, pp. 1291–1307. doi: 10.1109/TSMCC.2012.2218595.

Ho, Mark K., Fiery Cushman, Michael L. Littman, and Joseph L. Austerweil (2019). "People teach with rewards and punishments as communication, not reinforcements." In: *Journal of Experimental Psychology: General* 148.3, pp. 520–549. issn: 1939-2222(Electronic),0096-3445(Print). doi: 10.1037/xge0000569.

– (2021). "Communication in action: Planning and interpreting communicative demonstrations." In: *Journal of experimental psychology. General*.

Ho, Mark K., Michael L. Littman, Fiery Cushman, and Joseph L. Austerweil (2015). "Teaching with Rewards and Punishments: Reinforcement or Communication?" In: *CogSci*.

– (2018). "Effectively Learning from Pedagogical Demonstrations". In: *Annual Conference of the Cognitive Science Society (CogSci)*.

Ho, Mark K., Michael L. Littman, James MacGlashan, Fiery Cushman, and Joseph L. Austerweil (2016). "Showing versus doing: Teaching by demonstration". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., pp. 3027–3035.

Holladay, Rachel, Anca D. Dragan, and Siddhartha S. Srinivasa (Aug. 2014). "Legible Robot Pointing". In: *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*. Vol. 2014. doi: 10.1109/ROMAN.2014.6926256.

Huang, Sandy H., David Held, Pieter Abbeel, and Anca D. Dragan (2017). "Enabling Robots to Communicate Their Objectives". In: *ArXiv* abs/1702.03465.

Ijspeert, Auke Jan, J Nakanishi, and Stefan Schaal (2002). "Learning Attractor Landscapes for Learning Motor Primitives". In: *Advances in Neural Information Processing Systems 15 (NIPS2002)*, pp. 1547–1554.

Ijspeert, Auke Jan, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal (Nov. 2012). "Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors". In: *Neural computation* 25. doi: 10.1162/NECO_a_00393.

Ivaldi, Serena, Salvatore Anzalone, Woody Rousseau, Olivier Sigaud, and Mohamed Chetouani (2014). "Robot initiative in a team learning task increases the rhythm of interaction but not the perceived engagement". In: *Frontiers in Neurorobotics* 8, p. 5. issn: 1662-5218. doi: 10.3389/fnbot.2014.00005.

Johnson, Susan C. (2000). "The Recognition of Mentalistic Agents in Infancy". In: *Trends in Cognitive Sciences* 4.1, pp. 22–28. doi: 10.1016/s1364-6613(99)01414-x.

Kennedy, J., Paul Baxter, Emmanuel Senft, Tony Belpaeme, and S. Lemaignan (2021). "From Characterising Three Years of HRI to Methodology and Reporting Recommendations". In: vol. 2016-April, pp. 391–398. isbn: 978-1-4673-8370-7. doi: 10.1109/HRI.2016.7451777.

Khan, Faisal, Bilge Mutlu, and Xiaojin Zhu (2011). "How do humans teach: On curriculum learning and teaching dimension". In: *Advances in Neural Information Processing Systems*, pp. 1449–1457. ISSN: 9781618395993.

Knox, W. Bradley and Peter Stone (Aug. 2008). "TAMER: Training an Agent Manually via Evaluative Reinforcement". In: *2008 7th IEEE International Conference on Development and Learning*, pp. 292–297. DOI: 10.1109/DEVLRN.2008.4640845.

– (2009). "Interactively Shaping Agents via Human Reinforcement: The TAMER Framework". In: *Proceedings of the Fifth International Conference on Knowledge Capture*. K-CAP '09. Redondo Beach, California, USA: Association for Computing Machinery, pp. 9–16. ISBN: 9781605586588. DOI: 10.1145/1597735.1597738.

– (2010). "Combining Manual Feedback with Subsequent MDP Reward Signals for Reinforcement Learning". In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*. AAMAS '10. Toronto, Canada: International Foundation for Autonomous Agents and Multiagent Systems, pp. 5–12. ISBN: 9780982657119.

– (July 2011). "Augmenting Reinforcement Learning with Human Feedback". In: *ICML 2011 Workshop on New Developments in Imitation Learning*.

– (Sept. 2012a). "Reinforcement Learning from Human Reward: Discounting in Episodic Tasks". In: DOI: 10.1109/ROMAN.2012.6343862.

– (2012b). "Reinforcement Learning from Simultaneous Human and MDP Reward". In: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*. AAMAS '12. Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems, pp. 475–482. ISBN: 0981738117.

Knox, W. Bradley, Peter Stone, and Cynthia Breazeal (2013). "Training a Robot via Human Feedback: A Case Study". In: *Social Robotics*. Ed. by Guido Herrmann, Martin J. Pearson, Alexander Lenz, Paul Bremner, Adam Spiers, and Ute Leonards. Cham: Springer International Publishing, pp. 460–470. ISBN: 978-3-319-02675-6.

Kober, Jens, J. Andrew Bagnell, and Jan Peters (2013). "Reinforcement learning in robotics: A survey". In: *The International Journal of Robotics Research* 32.11, pp. 1238–1274. DOI: 10.1177/0278364913495721.

Kober, Jens, Betty Mohler, and Jan Peters (2008). "Learning perceptual coupling for motor primitives". In: *2008 IEEE/RSJ International Conference*

on Intelligent Robots and Systems, pp. 834–839. DOI: 10.1109/IROS.2008.4650953.

KOLTER, J. Zico and Andrew Y. NG (2009). "Policy search via the signed derivative". In: *Robotics: Science and Systems V, University of Washington, Seattle, USA, June 28 - July 1, 2009*. Ed. by Jeff TRINKLE, Yoky MATSUOKA, and José A. CASTELLANOS. The MIT Press. DOI: 10.15607/RSS.2009.V.027.

KORMUSHEV, Petar, Sylvain CALINON, and Darwin CALDWELL (Nov. 2010). "Robot Motor Skill Coordination with EM-based Reinforcement Learning". In: pp. 3232–3237. DOI: 10.1109/IROS.2010.5649089.

KUHLMANN, Gregory, Peter STONE, Raymond MOONEY, and Jude SHAVLIK (2004). "Guiding a reinforcement learner with natural language advice: Initial results in RoboCup soccer". In: *The AAAI-2004 workshop on supervisory control of learning and adaptive systems*. San Jose, CA.

KULKARNI, Anagha, Yantian ZHA, Tathagata CHAKRABORTI, Satya Gautam VADLAMUDI, Yu ZHANG, and Subbarao KAMBHAMPATI (2019). "Explicable Planning as Minimizing Distance from Expected Behavior". In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '19. Montreal QC, Canada: International Foundation for Autonomous Agents and Multiagent Systems, pp. 2075–2077. ISBN: 9781450363099.

LECLÈRE, C, M AVRIL, Sylvie VIAUX, N BODEAU, Catherine ACHARD, Sylvain MISSONNIER, Miri KEREN, Mohamed CHETOUANI, and Dana COHEN (May 2016). "Interaction and behaviour imaging: A novel method to measure mother-infant interaction using video 3D reconstruction". In: *Translational Psychiatry* 6. DOI: 10.1038/tp.2016.82.

LEYZBERG, Daniel, Samuel SPAULDING, Mariya TONEVA, and Brian SCASSELLATI (2012). "The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains". In: *34th Annual Conference of the Cognitive Science Society* 1, pp. 1882–1887. DOI: ISBN978-0-9768318-8-4.

LI, Guangliang, Randy GOMEZ, Keisuke NAKAMURA, and Bo HE (May 2019). "Human-Centered Reinforcement Learning: A Survey". In: *IEEE Transactions on Human-Machine Systems*. DOI: 10.1109/THMS.2019.2912447.

LIOUTIKOV, Rudolf, Gerhard NEUMANN, Guilherme MAEDA, and Jan PETERS (2015). "Probabilistic segmentation applied to an assembly task". In: *International Conference on Humanoid Robots*. ISSN: 21640580. DOI: 10.1109/HUMANOIDS.2015.7363584.

MacGlashan, James, Mark K. Ho, Robert Tyler Loftin, Bei Peng, David L. Roberts, Matthew E. Taylor, and Michael L. Littman (2017). "Interactive Learning from Policy-Dependent Human Feedback". In: *CoRR* abs/1701.06049.

Maclin, Richard, Jude Shavlik, Lisa Torrey, Trevor Walker, and Edward Wild (2005). "Giving Advice about Preferred Actions to Reinforcement Learners via Knowledge-Based Kernel Regression". In: *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*. AAAI'05. Pittsburgh, Pennsylvania: AAAI Press, pp. 819–824. isbn: 157735236x.

MacNally, Aleck M., Nir Lipovetzky, Miquel Ramirez, and Adrian R. Pearce (2018). "Action Selection for Transparent Planning". In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '18. Stockholm, Sweden: International Foundation for Autonomous Agents and Multiagent Systems, pp. 1327–1335.

Mainprice, J., R. Hayne, and D. Berenson (May 2015). "Predicting human reaching motion in collaborative tasks using Inverse Optimal Control and iterative re-planning". In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 885–892. doi: 10.1109/ICRA.2015.7139282.

Malle, B. and J. Knobe (1997). "The Folk Concept of Intentionality". In: *Journal of Experimental Social Psychology* 33, pp. 101–121.

Meltzoff, Andrew N. (1995). *Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children.* US. doi: 10.1037/0012-1649.31.5.838.

– (1999). "Born to Learn : What Infants Learn from Watching Us". In: Pediatric Institute Publications, pp. 1–10.

Milli, Smitha and Anca D. Dragan (2019). "Literal or Pedagogic Human? Analyzing Human Model Misspecification in Objective Learning". In: *CoRR* abs/1903.03877.

Moriarty, D. E., A. C. Schultz, and J. J. Grefenstette (Sept. 1999). "Evolutionary Algorithms for Reinforcement Learning". In: *Journal of Artificial Intelligence Research* 11, pp. 241–276. issn: 1076-9757. doi: 10.1613/jair.613.

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

Muelling, Katharina, Jens Kober, and Jan Peters (2010). "Learning table tennis with a Mixture of Motor Primitives". In: *2010 10th IEEE-RAS International Conference on Humanoid Robots*, pp. 411–416. DOI: 10.1109/ICHR.2010.5686298.

Mülling, Katharina, Jens Kober, Oliver Kroemer, and Jan Peters (2013). "Learning to select and generalize striking movements in robot table tennis". In: *The International Journal of Robotics Research* 32.3, pp. 263–279. DOI: 10.1177/0278364912472380.

Najar, Anis and Mohamed Chetouani (June 2021). "Reinforcement Learning With Human Advice: A Survey". In: *Frontiers in Robotics and AI*. DOI: 10.3389/frobt.2021.584075.

Najar, Anis, Olivier Sigaud, and Mohamed Chetouani (2019). "Interactively shaping robot behaviour with unlabeled human instructions". In: *CoRR* abs/1902.01670.

Nakanishi, Jun, Jun Morimoto, Gen Endo, Gordon Cheng, Stefan Schaal, and Mitsuo Kawato (2004). "Learning from demonstration and adaptation of biped locomotion". In: *Robotics and Autonomous Systems* 47.2. Robot Learning from Demonstration, pp. 79–91. ISSN: 0921-8890. DOI: https://doi.org/10.1016/j.robot.2004.03.003.

Nehaniv, Chrystopher L. and Kerstin Dautenhahn (1999). "Of Hummingbirds and Helicopters: an Algebraic Framework for Interdisciplinary Studies of Imitation and Its Applications". In: *Interdisciplinary Approaches to Robot Learning*.

Ng, Andrew Y., Daishi Harada, and Stuart J. Russell (1999). "Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping". In: *Proceedings of the Sixteenth International Conference on Machine Learning*. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 278–287. ISBN: 1558606122.

Ng, Andrew Y. and Michael Jordan (2000). "PEGASUS: A policy search method for large MDPs and POMDPs". In: *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence (UAI 2000)*. Ed. by Craig Boutilier and Moisés Goldszmidt. San Francisco, CA, USA: Morgan Kaufmann, pp. 406–415. ISBN: 1-55860-709-9.

Nomikou, Iris, Giuseppe Leonardi, Katharina Rohlfing, and Joanna Rączaszek-Leonardi (May 2016). "Constructing Interaction: The De-

velopment of Gaze Dynamics: Development of Gaze Dynamics in Interaction". In: *Infant and Child Development* 25. DOI: 10.1002/icd.1975.

OSA, Takayuki, Joni PAJARINEN, Gerhard NEUMANN, J. Andrew BAGNELL, Pieter ABBEEL, and Jan PETERS (2018). "An Algorithmic Perspective on Imitation Learning". In: *CoRR* abs/1811.06711.

PAIS URECHE, Ana-Lucia and Aude BILLARD (2015). "Learning Bimanual Coordinated Tasks From Human Demonstrations". In: *Human-Robot Interaction. HRI'15 Extended Abstracts*. DOI: 10.1145/2701973.2702007.

PALÉOLOGUE, Victor, Jocelyn MARTIN, Amit Kumar PANDEY, and Mohamed CHETOUANI (2018). "Semantic-Based Interaction for Teaching Robot Behavior Compositions Using Spoken Language". In: *Social Robotics*. Ed. by Shuzhi Sam GE, John-John CABIBIHAN, Miguel A. SALICHS, Elizabeth BROADBENT, Hongsheng HE, Alan R. WAGNER, and Álvaro CASTRO-GONZÁLEZ. Cham: Springer International Publishing, pp. 421–430. ISBN: 978-3-030-05204-1.

PARASCHOS, Alexandros, Christian DANIEL, Jan PETERS, and Gerhard NEUMANN (2013). "Probabilistic Movement Primitives". In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI, and K. Q. WEINBERGER. Vol. 26. Curran Associates, Inc.

– (2018). "Using probabilistic movement primitives in robotics". In: *Autonomous Robots* 42.3, pp. 529–551. ISSN: 15737527. DOI: 10.1007/s10514-017-9648-7.

PEZZULO, Giovanni, Francesco DONNARUMMA, and Haris DINDO (2013). "Human sensorimotor communication: A theory of signaling in online social interactions". In: *PLoS ONE*. ISSN: 19326203. DOI: 10.1371/journal.pone.0079876.

PEZZULO, Giovanni, Francesco DONNARUMMA, Haris DINDO, Alessandro D'AUSILIO, Ivana KONVALINKA, and Cristiano CASTELFRANCHI (2019). "The body talks: Sensorimotor communication and its brain and kinematic signatures". In: *Physics of Life Reviews* 28, pp. 1–21. ISSN: 15710645. DOI: 10.1016/j.plrev.2018.06.014.

PREMACK, David and G. WOODRUFF (1978). "Does the Chimpanzee Have a Theory of Mind?" In: *Behavioral and Brain Sciences* 4.4, pp. 515–629. DOI: 10.1017/S0140525X00076512.

QI, S. and S. ZHU (May 2018). "Intent-Aware Multi-Agent Reinforcement Learning". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7533–7540. DOI: 10.1109/ICRA.2018.8463211.

Reddy, Siddharth, Anca D. Dragan, Sergey Levine, Shane Legg, and Jan Leike (2021). *Learning Human Objectives by Evaluating Hypothetical Behavior*.

Rosenstein, Michael and Andrew G. Barto (Dec. 2003). "1 Supervised Actor-Critic Reinforcement Learning". In: *ACM Sigevolution*.

Sciutti, Alessandra, Caterina Ansuini, Cristina Becchio, and Giulio Sandini (Sept. 2015). "Investigating the ability to read others' intentions using humanoid robots". In: *Frontiers in Psychology* 6. DOI: 10.3389/fpsyg.2015.01362.

Searle, John R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press. DOI: 10.1017/CBO9781139173438.

Sena, A., Y. Zhao, and M. J. Howard (May 2018). "Teaching Human Teachers to Teach Robot Learners". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–7. DOI: 10.1109/ICRA.2018.8461194.

Shadmehr, Reza, Maurice Smith, and John Krakauer (Mar. 2010). "Error Correction, Sensory Prediction, and Adaptation in Motor Control". In: *Annual review of neuroscience* 33, pp. 89–108. DOI: 10.1146/annurev-neuro-060909-153135.

Shafto, Patrick, Noah D. Goodman, and Michael C. Frank (2012). "Learning From Others: The Consequences of Psychological Reasoning for Human Learning". In: *Perspectives on Psychological Science* 7.4. PMID: 26168471, pp. 341–351. DOI: 10.1177/1745691612448481.

Shafto, Patrick, Noah D. Goodman, and Thomas L. Griffiths (2014). "A rational account of pedagogical reasoning: Teaching by, and learning from, examples". In: *Cognitive Psychology* 71, pp. 55–89.

Shannon, Claude E. and Warren Weaver (1949). *The Mathematical Theory of Communication*. Urbana and Chicago: University of Illinois Press.

Sigaud, Olivier, Hugo Caselles-Dupré, Cédric Colas, Ahmed Akakzia, Pierre-Yves Oudeyer, and Mohamed Chetouani (2021). "Towards Teachable Autonomous Agents". In: *CoRR* abs/2105.11977.

Sperber, Dan and Deirdre Wilson (1995). *Relevance: Communication and cognition, 2nd ed.* Malden: Blackwell Publishing, pp. viii, 326–viii, 326. ISBN: 0-631-19878-4 (Paperback).

Steffen, Jan, Christof Elbrechter, Robert Haschke, and Helge Ritter (2010). "Bio-inspired motion strategies for a bimanual manipulation task".

In: *International Conference on Humanoid Robots*. ISSN: 2164-0572. DOI: 10.1109/ICHR.2010.5686830.

SUAY, Halit Bener and Sonia CHERNOVA (2011). "Effect of human guidance and state space size on Interactive Reinforcement Learning". In: *2011 RO-MAN*, pp. 1–6. DOI: 10.1109/ROMAN.2011.6005223.

SUBRAMANIAN, Kaushik, Charles L. ISBELL, and Andrea L. THOMAZ (2016). "Exploration from Demonstration for Interactive Reinforcement Learning". In: *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. AAMAS '16. Singapore, Singapore: International Foundation for Autonomous Agents and Multiagent Systems, pp. 447–456. ISBN: 9781450342391.

SUTTON, Richard S. (1996). "Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding". In: *Advances in Neural Information Processing Systems*. Ed. by D. TOURETZKY, M. C. MOZER, and M. HASSELMO. Vol. 8. MIT Press.

SUTTON, Richard S. and Andrew G. BARTO (1998). *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA: MIT Press. ISBN: 0262193981.

SUTTON, Richard S., David MCALLESTER, Satinder SINGH, and Yishay MANSOUR (Feb. 2000). "Policy Gradient Methods for Reinforcement Learning with Function Approximation". In: *Adv. Neural Inf. Process. Syst* 12.

THOMAZ, Andrea L. and Cynthia BREAZEAL (2006a). "Reinforcement Learning with Human Teachers: Evidence of Feedback and Guidance with Implications for Learning Performance". In: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*. AAAI'06. Boston, Massachusetts: AAAI Press, pp. 1000–1005. ISBN: 9781577352815.

– (2006b). "Transparency and Socially Guided Machine Learning". In: *5th Intl. Conf. on Development and Learning (ICDL)*.

– (2008). "Teachable robots: Understanding human teaching behavior to build more effective robot learners". In: *Artificial Intelligence*. ISSN: 00043702. DOI: 10.1016/j.artint.2007.09.009.

THORNDIKE, Edward L. (1898). "Animal intelligence: An experimental study of the associative processes in animals." In: *The Psychological Review: Monograph Supplements* 2.4, pp. i–109. ISSN: 0096-9753(Print). DOI: 10.1037/h0092987.

TOMASELLO, Michael, Malinda CARPENTER, Josep CALL, Tanya BEHNE, and Henrike MOLL (2005). "Understanding and sharing intentions: The origins

of cultural cognition". In: *Behavioral and Brain Sciences* 28.5, pp. 675–691. DOI: 10.1017/S0140525X05000129.

TORREY, Lisa, Jude SHAVLIK, Trevor WALKER, and Richard MACLIN (2006). "Skill Acquisition Via Transfer Learning and Advice Taking". In: *Machine Learning: ECML 2006*. Ed. by Johannes FÜRNKRANZ, Tobias SCHEFFER, and Myra SPILIOPOULOU. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 425–436. ISBN: 978-3-540-46056-5.

TURING, A. M. (Oct. 1950). "I.—COMPUTING MACHINERY AND INTELLIGENCE". In: *Mind* LIX.236, pp. 433–460. ISSN: 0026-4423. DOI: 10.1093/mind/LIX.236.433.

VARNI, James, O. LOVAAS, Robert KOEGEL, and Nancy EVERETT (Apr. 1979). "An analysis of observational learning in autistic and normal children". In: *Journal of abnormal child psychology* 7, pp. 31–43. DOI: 10.1007/BF00924508.

VIEN, Ngo Anh, Wolfgang ERTEL, and Tae Choong CHUNG (2013). "Learning via human feedback in continuous state and action spaces". In: *Applied Intelligence* 39.2, pp. 267–278. ISSN: 1573-7497. DOI: 10.1007/s10489-012-0412-6.

VOLLMER, Anna-Lisa and Lars SCHILLINGMANN (2018). "On Studying Human Teaching Behavior with Robots: A Review". In: *Review of Philosophy and Psychology* 9.4, pp. 863–903. DOI: 10.1007/s13164-017-0353-4.

WALLKOTTER, Sebastian, Silvia TULLI, Ginevra CASTELLANO, Ana PAIVA, and Mohamed CHETOUANI (2020). *Explainable Agents Through Social Cues: A Review*.

WATKINS, Christopher J. C. H. and Peter DAYAN (1992). "Q-learning". In: *Machine Learning*, pp. 279–292.

WOLPERT, Daniel M and Zoubin GHAHRAMANI (2000). *Computational principles of movement neuroscience.* United Kingdom. DOI: 10.1038/81497.

WYKOWSKA, Agnieszka, Thierry CHAMINADE, and Gordon CHENG (2016). "Embodied artificial agents for understanding human social cognition". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1693, p. 20150375. DOI: 10.1098/rstb.2015.0375.

ZHANG, Yu, Sarath SREEDHARAN, Anagha KULKARNI, Tathagata CHAKRABORTI, Hankz Hankui ZHUO, and Subbarao KAMBHAMPATI (July 2017). "Plan explicability and predictability for robot task planning". English (US). In: *ICRA 2017 - IEEE International Conference on Robotics and Automation*. United

States: Institute of Electrical and Electronics Engineers Inc., pp. 1313–1320. DOI: 10.1109/ICRA.2017.7989155.

Zhu, Xiaojin, Adish Singla, Sandra Zilles, and Anna N. Rafferty (2018). "An Overview of Machine Teaching". In: *CoRR* abs/1801.05927.

Ziebart, Brian D., Andrew Maas, J. Andrew Bagnell, and Anind K. Dey (Jan. 2008). "Maximum Entropy Inverse Reinforcement Learning". In: *Proc. AAAI*, pp. 1433–1438.

Appendix

A User Study Forms

INSEAD UNIVERSITY
France Campus: Fontainebleau Cedex, 77305
Singapore Campus: 1 Ayer Rajah Avenue, 138676

Craig Smith, Ph. D. +33 (0)1 60 72 00 00
CHAIR, PANEL ON HUMAN PARTICIPANTS RESEARCH

NOTICE ON FULL ETHICAL REVIEW

Date: March 19, 2019

To: Mohamed Chetouani & Manuel Bied

From: Craig Smith, PhD, Chair, Institutional Review Board on Human Participants Research

Protocol: Human Teaching Behavior Towards Robots

Protocol ID: March 2019/1

The INSEAD ethical review committee has reviewed your research protocol on 15 March 2019 and determined that the study is approved without changes. If this protocol is used in conjunction with any other human use, it must be re-reviewed. The ethical committee requests prompt notification of any complications or incidents of noncompliance, which may occur during any human use procedure.

Please remember that in case any new data is collected, all data including the consent forms must be retained for a minimum of three years past the completion of this research. Additional requirement may be imposed by your funding agency, your department, or other entities.

Yours sincerely,

Craig Smith
Review type: FULL – NEW

Europe Campus – Boulevard de Constance
77305 Fontainebleau Cedex, France
Tel: +33 (0)1 60 72 40 00

www.insead.edu

Figure A.1: IRB approval letter of the user study

FEUILLE DE CONSENTEMENT

Etude SU - Apprentissage dans une interaction Humain-Robot

Bienvenue au Centre Multidisciplinaire des Sciences Comportementales Sorbonne Université-INSEAD. Les chercheurs de l'étude à laquelle vous allez participer sont :
- Mohamed Chetouani (Professeur à Sorbonne université)
- Manuel Bied (Doctorant à Sorbonne université)

Dans cette étude, vous donnerez des exemples sur la façon de résoudre des labyrinthes en 2 dimensions sur une tablette. Les exemples doivent être donnés de manière à ce qu'un robot puisse apprendre à partir de ces exemples comment résoudre ces labyrinthes. Votre participation durera environ 40 minutes.

Si vous terminez cette étude, vous recevrez 8 euros.

Nos études sont à visées académiques, et les résultats seront accessibles dans des publications scientifiques. Nous ne réalisons pas d'études pour le compte d'entreprises privées. Il n'existe aucun risque lié à cette étude autre que ceux de la vie de tous les jours.

Par ailleurs, vous pourrez recevoir des informations sur les conclusions de l'étude si vous le souhaitez, ainsi que des références concernant le type de recherche auquel vous avez participé. Cependant, parce que vos réponses sont anonymes, nous ne pourrons vous renseigner que sur les résultats agrégés de l'étude, et non sur votre performance ou les performances de n'importe qui d'autre ayant participé.

Les données concernant cette étude seront conservées sous clé ou protégées par un mot de passe, et seront détruites dès lors qu'elles ne seront plus utilisées.

Votre participation à l'étude doit être entièrement volontaire, et vous avez la possibilité de vous retirer de l'étude à tout moment sans aucune pénalité.

Je déclare être majeur(e), et ayant lu et parfaitement compris les paragraphes ci-dessus, accepte de mon plein gré de participer à cette étude.

DATE :

NOM :

PRÉNOM :

SIGNATURE :

Figure A.2: Consent form of the first experiment of the user study.

INSEAD Centre Multidisciplinaire des Sciences Comportementales Sorbonne Université-INSEAD SORBONNE UNIVERSITÉ CRÉATEURS DE FUTURS DEPUIS 1257

DEBRIEFING

Étude « Apprentissage dans une interaction Humain-Robot »

Explication de ce type de recherche

La programmation des robots demande souvent beaucoup de temps et cela ne leur permet que de résoudre des tâches spécifiques. De ce fait, nous nous intéressons à la capacité qu'on les robots d'apprendre des gens. La plupart des approches précédentes portaient principalement sur le robot et son apprentissage et peu de recherches se sont intéressé aux aspects relatifs à l'humain. Afin de déployer des robots dans des environnements de la vie quotidienne, il est important de comprendre comment les humains interagissent avec des robots ou des agents virtuels afin d'intégrer leurs besoins et de prendre en compte leurs comportements.

Description plus détaillée de l'étude

L'étude d'aujourd'hui visait à déterminer comment les gens donnent des exemples de solutions pour résoudre un labyrinthe. Nous nous attendons à ce que les gens aient différentes stratégies d'enseignement qui, pour certaines, pourraient conduire à de meilleurs résultats d'apprentissage que d'autres. Nous nous attendons à ce que des participants fassent de meilleures démonstrations après avoir deviné quelle est la position du terrain dangereux.
Les données collectées seront utilisées pour évaluer le fonctionnement d'algorithmes de pointe pour identifier d'éventuels défauts de ces algorithmes et identifier l'impact que peut avoir le fait que des personnes enseignent à un robot sans avoir reçu de formation sur l'enseignement à un robot ou à un agent virtuel.

Vous pouvez trouver plus d'informations concernant ce type d'études à l'aide de la référence ci-dessous :

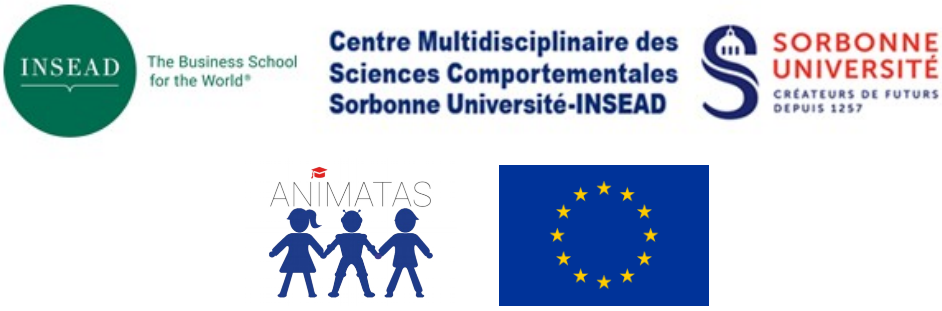
- Khan, Mutlu, & Zhu. "How do humans teach: On curriculum learning and teaching dimension". In: Advances in Neural Information Processing Systems (2011), pp. 1449–145.
- Cakmak & Thomaz. "Eliciting good teaching from humans for machine learners". In: Artificial Intelligence 217 (2014), pp. 198–215

Si vous avez des questions concernant cette étude, vous pouvez contacter :
Mohamed Chetouani : mohamed.chetouani@sorbonne-universite.fr

Courte présentation du chercheur

Mohamed Chetouani est à la tête de l'équipe IMI2S (interaction, intégration multimodale et Social Signal) à l'Institut de Systèmes Intelligents et de Robotique (CNRS UMR 7222), Sorbonne Université. Il est actuellement professeur titulaire en traitement du signal, reconnaissance des formes et en machine-learning.

Figure A.3: Debriefing form of the first experiment of the user study.



FEUILLE DE CONSENTEMENT

Etude SU - Apprentissage dans une interaction Humain-Robot 2

Bienvenue au Centre Multidisciplinaire des Sciences Comportementales Sorbonne Université-INSEAD. Les chercheurs de l'étude sont Mohamed Chetouani (mohamed.chetouani@sorbonne-universite.fr) et Manuel Bied (bied.manuel@gmail.com)

Dans cette étude, il vous sera montré des démonstrations que des participants précédents ont donné à un robot pour résoudre une tâche dans un labyrinthe à deux dimensions. Il vous sera demandé d'évaluer la simplicité de ces démonstrations et s'il y a des informations supplémentaires dans la démonstration. Votre participation durera approximativement 60 minutes.

Nos études sont à visées académiques, et les résultats seront accessibles dans des publications scientifiques. Nous ne réalisons pas d'études pour le compte d'entreprises privées. Il n'existe aucun risque lié à cette étude autre que ceux de la vie de tous les jours.

Si vous terminez cette étude, vous recevrez 10 euros.

Par ailleurs, vous pourrez recevoir des informations sur les conclusions de l'étude si vous le souhaitez, ainsi que des références concernant le type de recherche auquel vous avez participé. Cependant, parce que vos réponses sont anonymes, nous ne pourrons vous renseigner que sur les résultats agrégés de l'étude, et non sur votre performance ou les performances de n'importe qui d'autre ayant participé.

Les données concernant cette étude seront conservées sous clé ou protégées par un mot de passe, et seront détruites dès lors qu'elles ne seront plus utilisées.

Votre participation à l'étude doit être entièrement volontaire, et vous avez la possibilité de vous retirer de l'étude à tout moment sans aucune pénalité.

Je déclare être majeur(e), et ayant lu et parfaitement compris les paragraphes ci-dessus, accepte de mon plein gré de participer à cette étude.

DATE :

NOM :

PRÉNOM :

Figure A.4: Consent form of the second experiment of the user study.

Explication du domaine de recherche général

Programmer des robots prend beaucoup de temps et leur permet seulement de résoudre des tâches spécifiques. Par conséquent, il est important de les enrichir avec la capacité d'apprendre des autres. La plupart des approches précédentes se concentrent surtout sur le robot et comment il peut apprendre. Le côté de l'homme est souvent négligé. Pour pouvoir déployer des robots dans des environnements de la vie de tous les jours, il est important de comprendre comment les humains interagissent avec les robots ou les agents virtuels pour intégrer leurs besoins et représenter leurs comportements. Un domaine spécial d'intérêt est l'apprentissage et l'enseignement de compétences sensorimotrices. Les hommes peuvent exagérer leurs mouvements pour signaler leurs intentions aux autres humains. Les mécanismes intervenants dans l'interaction homme-robot ont besoin d'être explorés, en particulier la façon d'apprendre des compétences aux robots.

Description détaillée de l'étude

L'étude d'aujourd'hui a besoin d'être vue dans le contexte de l'étude de l'utilisateur précédent, où l'on a demandé aux personnes d'enseigner comment résoudre un labyrinthe à un robot. De plus, on leur a demandé de résoudre seulement la tâche.

De sorte à avoir plus d'idées à propos des démonstrations, l'étude d'aujourd'hui a été menée pour évaluer les perceptions des individus sur celles-ci. Ces perceptions seront utilisées pour comprendre les différences des démonstrations en terme de qualité de l'enseignement, de qualité de résolution et pour permettre la création de 'clusters' de différents types de démonstrations.
Pour finir, les informations collectées sur les études précédentes et sur cette étude devraient être utilisées pour comprendre les comportements d'enseignement envers des robots et pour permettre la construction d'algorithmes qui 'comprennent' différentes intensions d'enseignement des hommes.

Vous pouvez trouver plus d'informations concernant l'étude ci-dessous:

- *Faisal Khan, Bilge Mutlu, and Xiaojin Zhu. "How do humans teach: On curriculum learning and teaching dimension". In:Advances in Neural Information Processing Systems (2011), pp. 1449–145*
- *Maya Cakmak and Andrea L. Thomaz. "Eliciting good teaching from humans for machine learners". In: Artificial Intelligence 217 (2014), pp. 198–215*
- ***Human Sensorimotor Communication: A Theory of Signaling in Online Social Interactions*** *Pezzulo G, Donnarumma F, Dindo H (2013) Human Sensorimotor Communication: A Theory of Signaling in Online Social Interactions. PLOS ONE 8(11): e79876.*

Si vous avez plus de questions, vous pouvez contacter :
Manuel BIED (manuel.bied@isir.upmc.fr), Mohamed CHETOUANI (mohamed.chetouani@sorbonne-universite.fr)

Figure A.5: Debriefing form of the second experiment of the user study (page 1).

Le chercheur
Mohamed Chetouani est directeur de l'équipe IMI2S (Interaction, intégration multimodale et Social Signal) à l'Institut de Systèmes Intelligents et de Robotique (CNRS UMR 7222), Sorbonne Université. Il est actuellement professeur en traitement du signal, reconnaissance des formes et machine-learning.

Figure A.6: Debriefing form of the second experiment of the user study (page 2).

Appendix B Environments Experiment 1 (User Study)

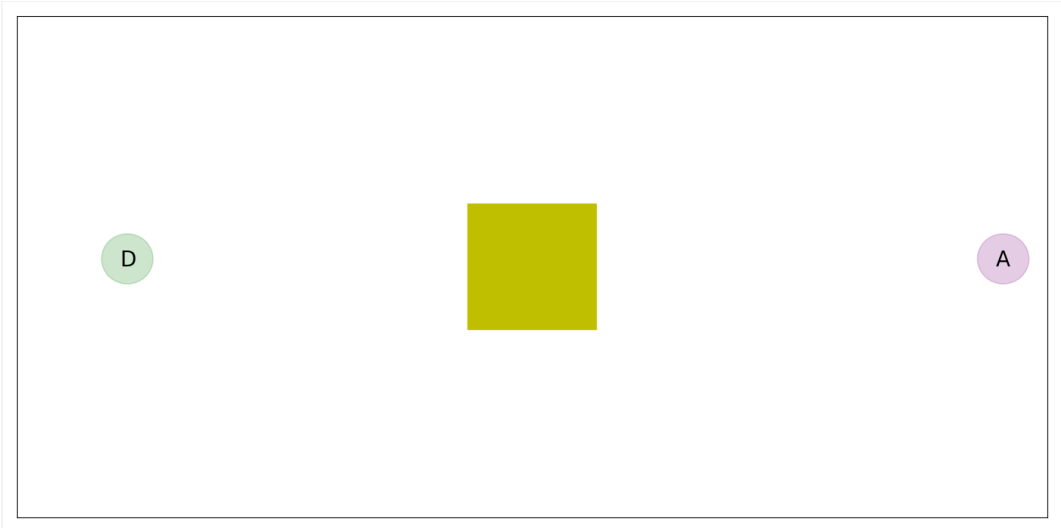


Figure B.1: Environment 1 of the user study.

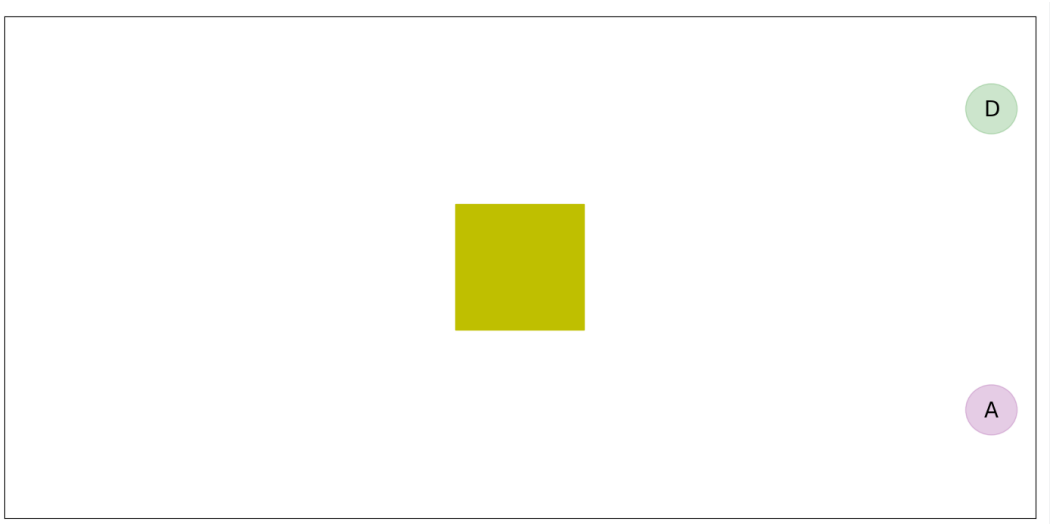


Figure B.2: Environment 2 of the user study.

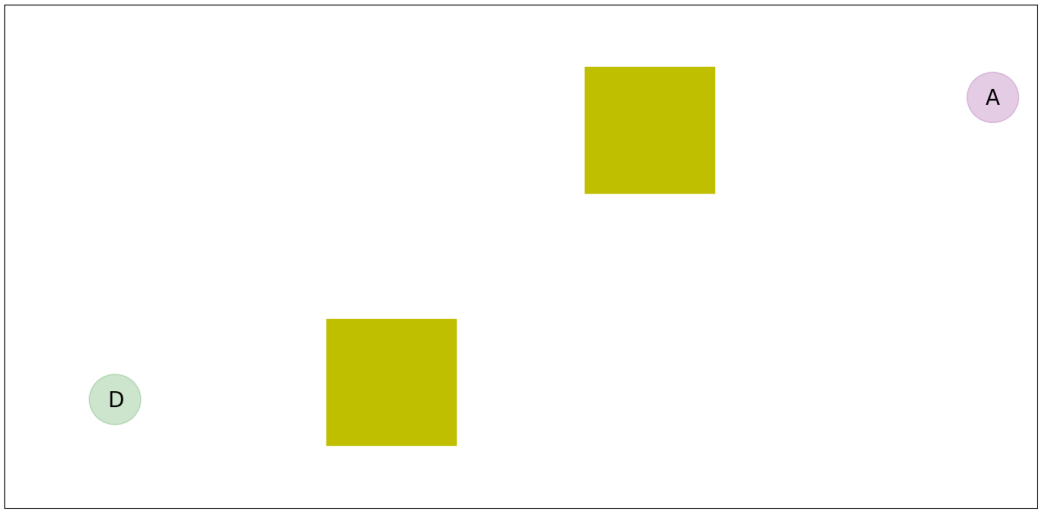


Figure B.3: Environment 3 of the user study.

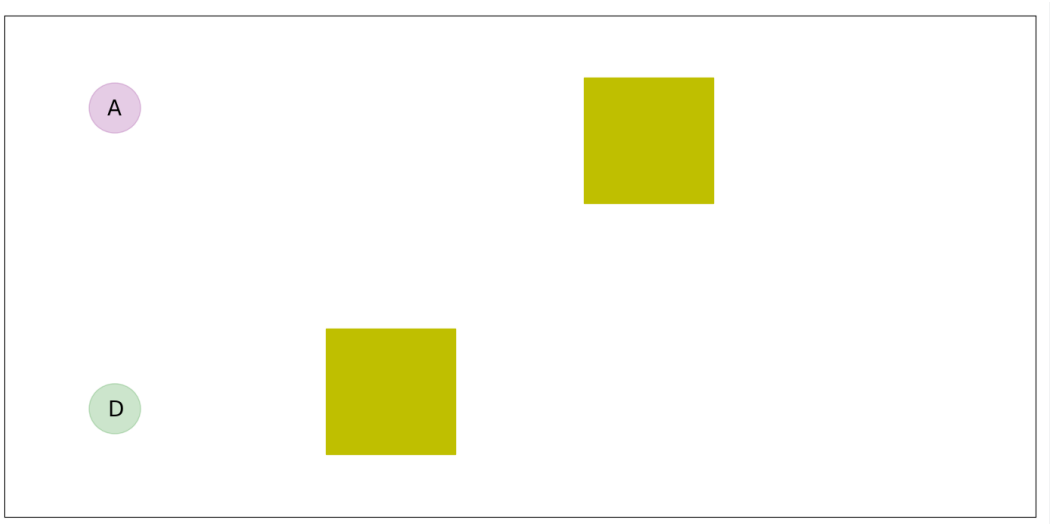


Figure B.4: Environment 4 of the user study.

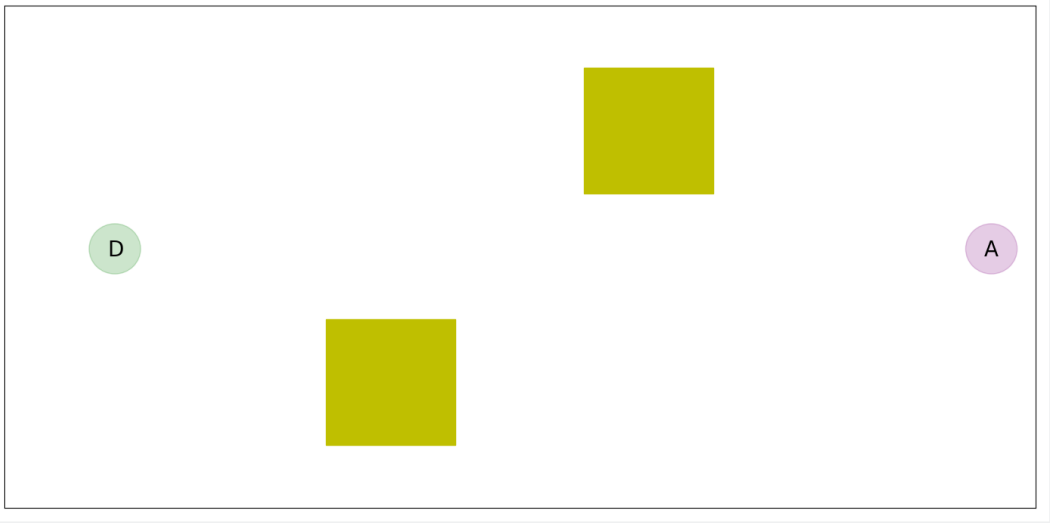


Figure B.5: Environment 5 of the user study.

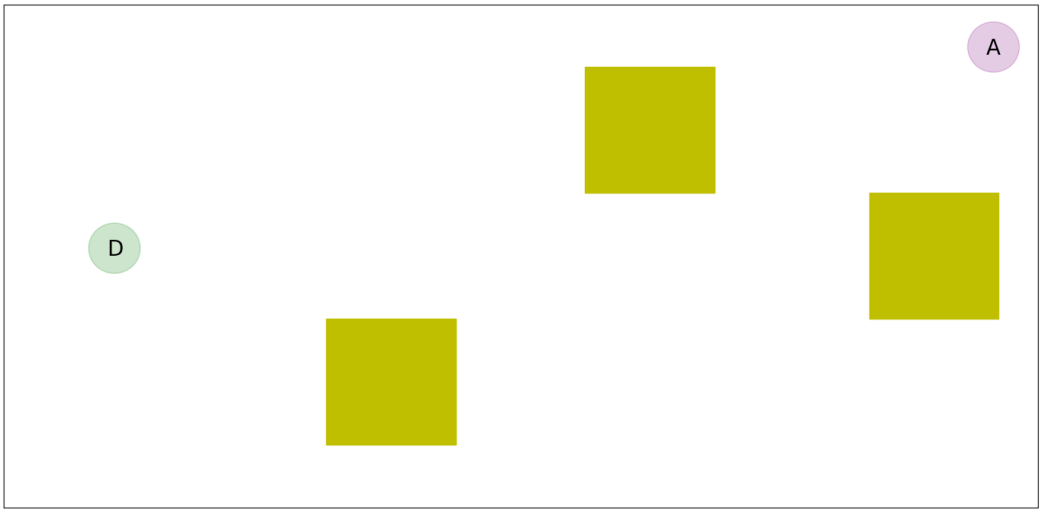


Figure B.6: Environment 6 of the user study.

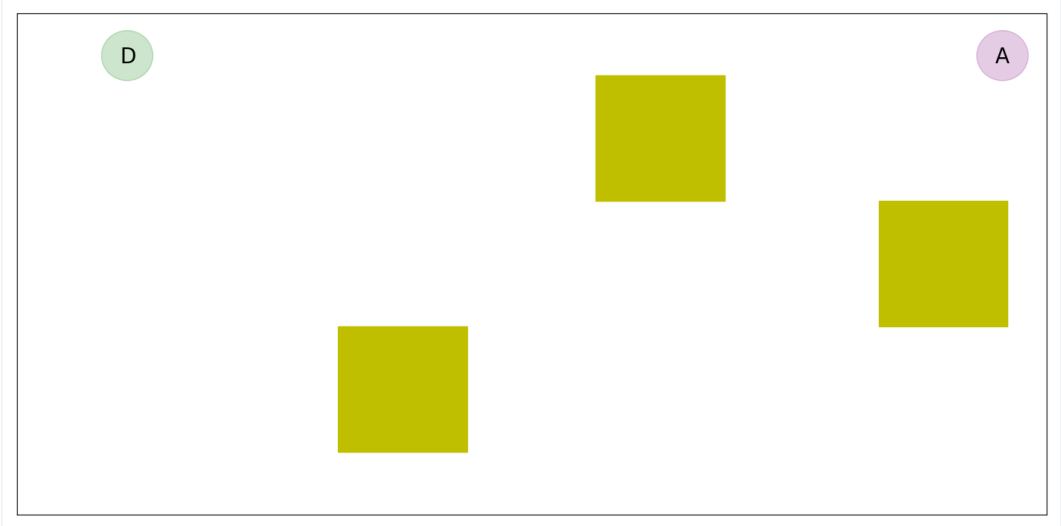


Figure B.7: Environment 7 of the user study.

Figure B.8: Environment 8 of the user study.

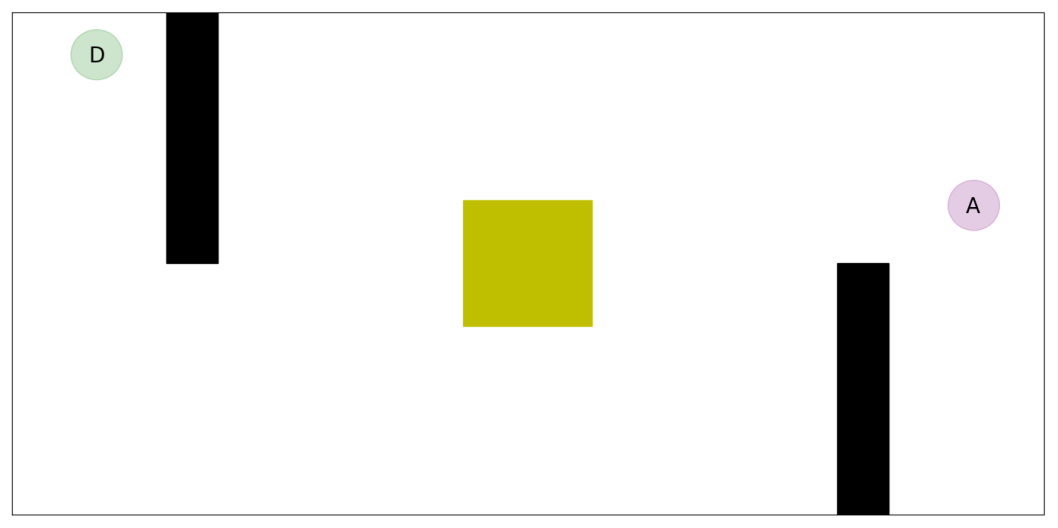


Figure B.9: Environment 9 of the user study.

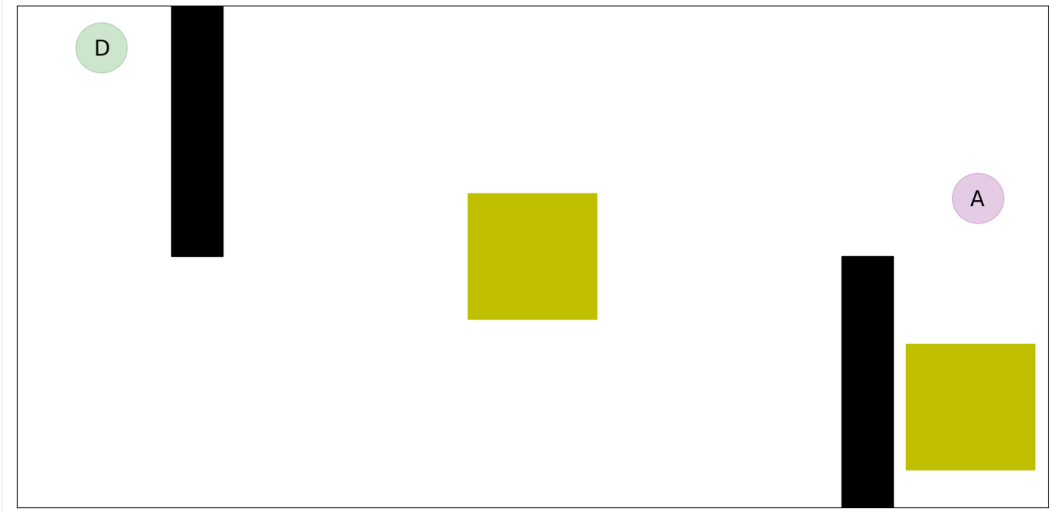


Figure B.10: Environment 10 of the user study.

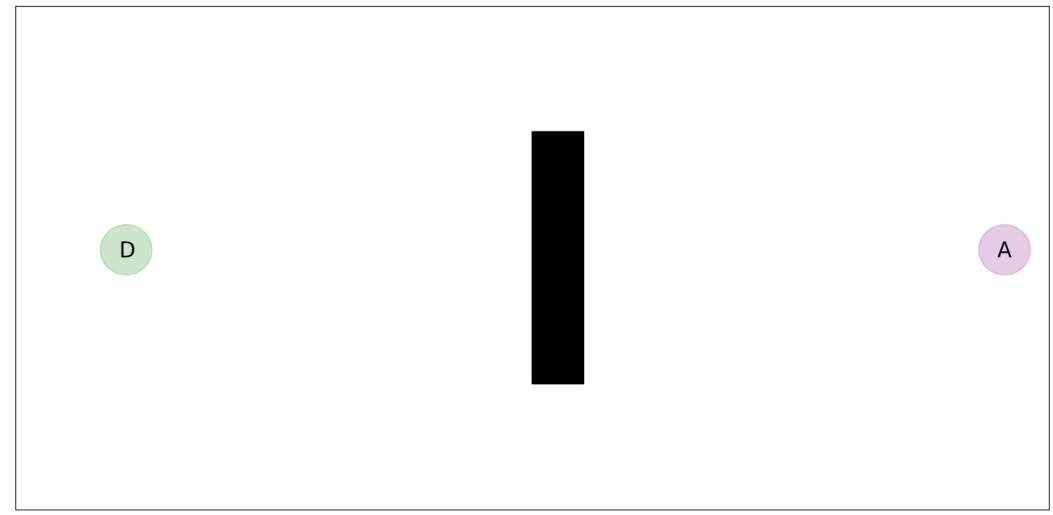


Figure B.11: Environment 11 of the user study.

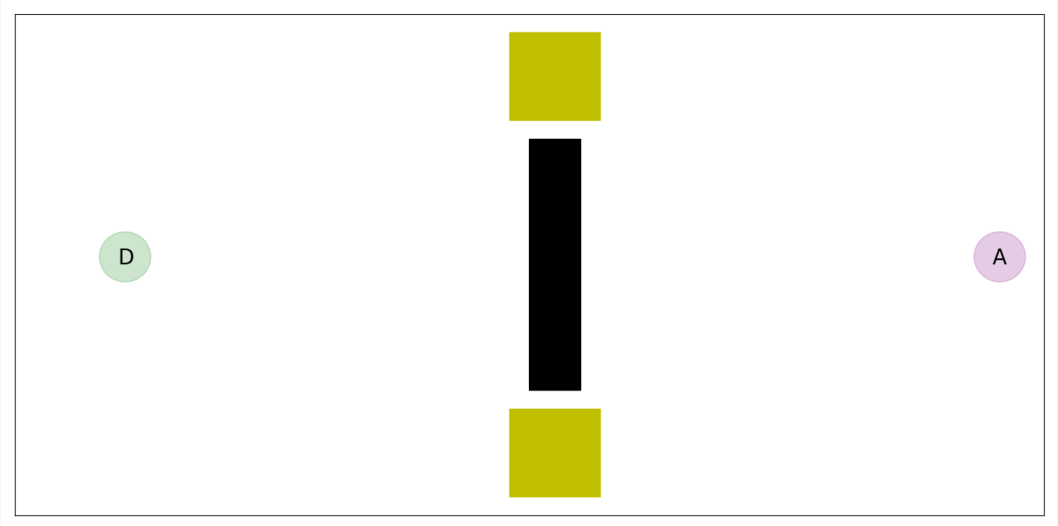


Figure B.12: Environment 12 of the user study.

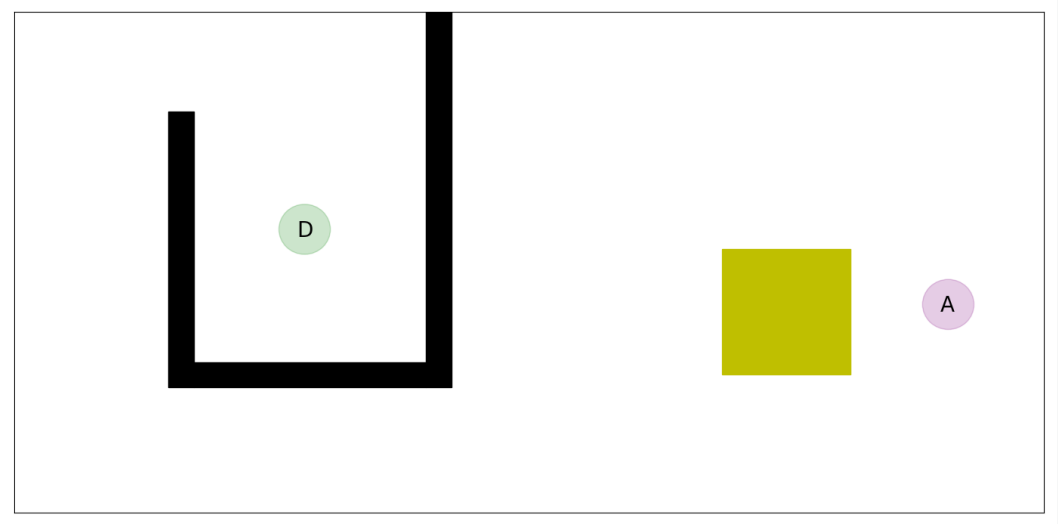


Figure B.13: Environment 13 of the user study.

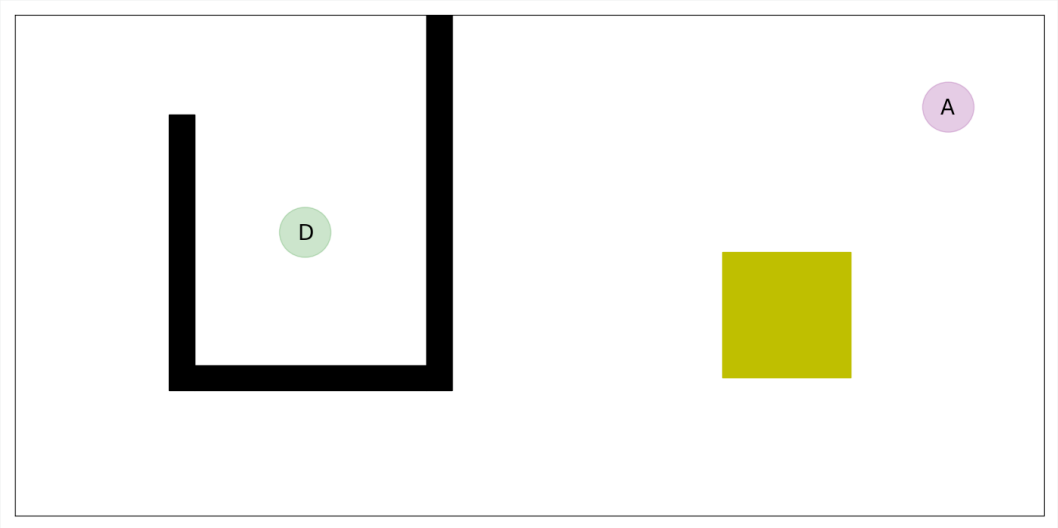


Figure B.14: Environment 14 of the user study.

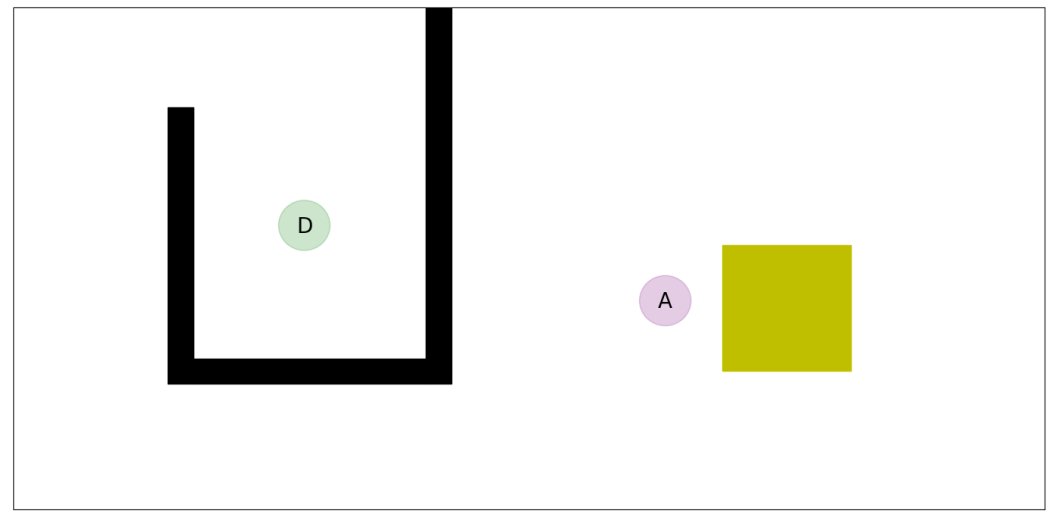


Figure B.15: Environment 15 of the user study.

Appendix

C GUI of Experiment 2 (User Study)

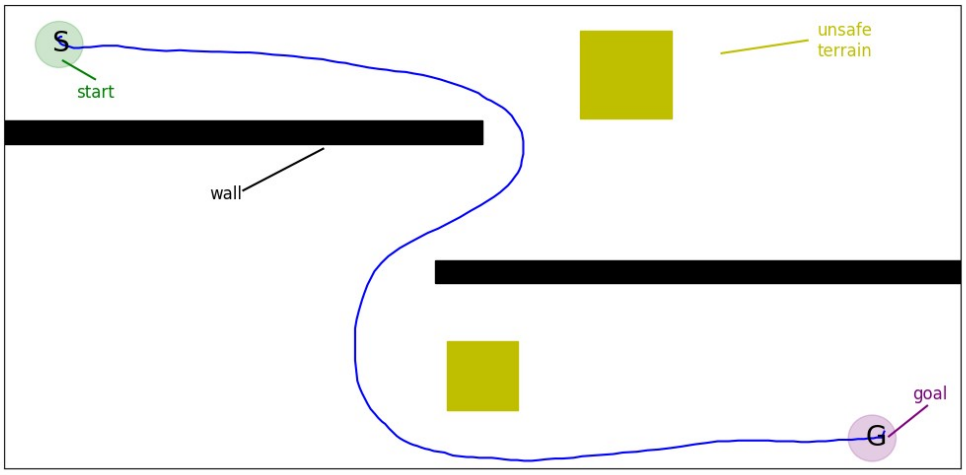
Description of the following steps:

In this user study you will be asked to give your impression about the quality of demonstrations of how to solve a certain task.These demonstrations were collected from a previous user study where people were asked to demonstrate how to solve this task to a robot. On the next page, we will describe the task that people were asked to do. Subsequently, we explain how the demonstration process worked. You will not need to solve the task, but only to rate demonstrations according to a given rating scheme.

Continue

Description of the task

The task is to solve a maze like shown below from the start zone (S) to the goal zone (G).There are unsafe terrains that could be entered, but it is not safe because the robot could get stuck or even break in this area.The walls are impassable and can't be crossed.



Start (S): The demonstrated trajectories must start within the start circle.

wall: This area can't be crossed.

Goal (G): The demonstrated trajectories must end within the goal circle.

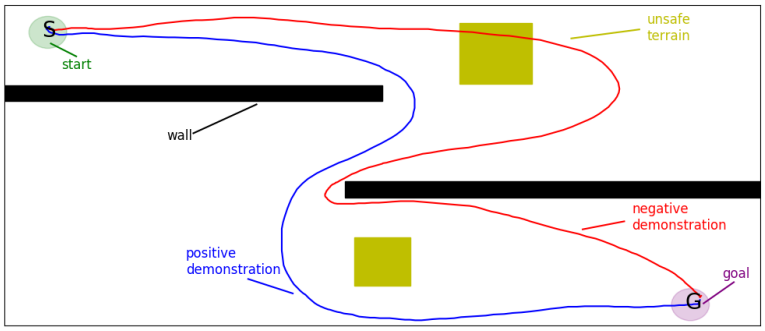
Unsafe terrain: This terrain shouldn't be crossed.

Continue

Figure C.1: GUI of experiment 2 (part 1).

Description of the teaching process:

In order to demonstrate to the robot how it could solve the maze, people could give positive and negative demonstrations. Positive demonstrations are correct solutions of the task, negative demonstrations are solutions that go through the unsafe terrain. The people were told that after the teaching phase the robot should learn from the given examples how to generate paths through the maze on its own. Furthermore, the information was given that they can assume that the robot knows the location of the walls but not of the unsafe terrain.



Start (S): The demonstrated trajectories must start within the start circle. **Unsafe terrain:** This terrain shouldn't be crossed. The robot doesn't know the location.

Goal (G): The demonstrated trajectories must end within the goal circle. **Positive demonstration:** An example for a trajectory that does not cross the unsafe terrain.

wall: This area can't be crossed. The robot knows the location. **Negative demonstration:** An example for a trajectory that cross the unsafe terrain.

Continue

Explanation of the rating scheme:

We will show you different demonstrations and you will be asked to rate each demonstration on a scale with five levels from strongly disagree to strongly agree.

The demonstrator is solving the task in the simplest way possible

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

In addition to solving the task, the demonstrator tries to convey you information about it

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I'm confident with my ratings

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

Reminder: 'solving the task' refers to going from the starting to the goal zone WITHOUT going through the unsafe terrain for positive demonstrations (blue) and going from the start to the goal zone AND going through the unsafe terrain for negative demonstrations (red).

Continue

Figure C.2: GUI of experiment 2 (part 2).

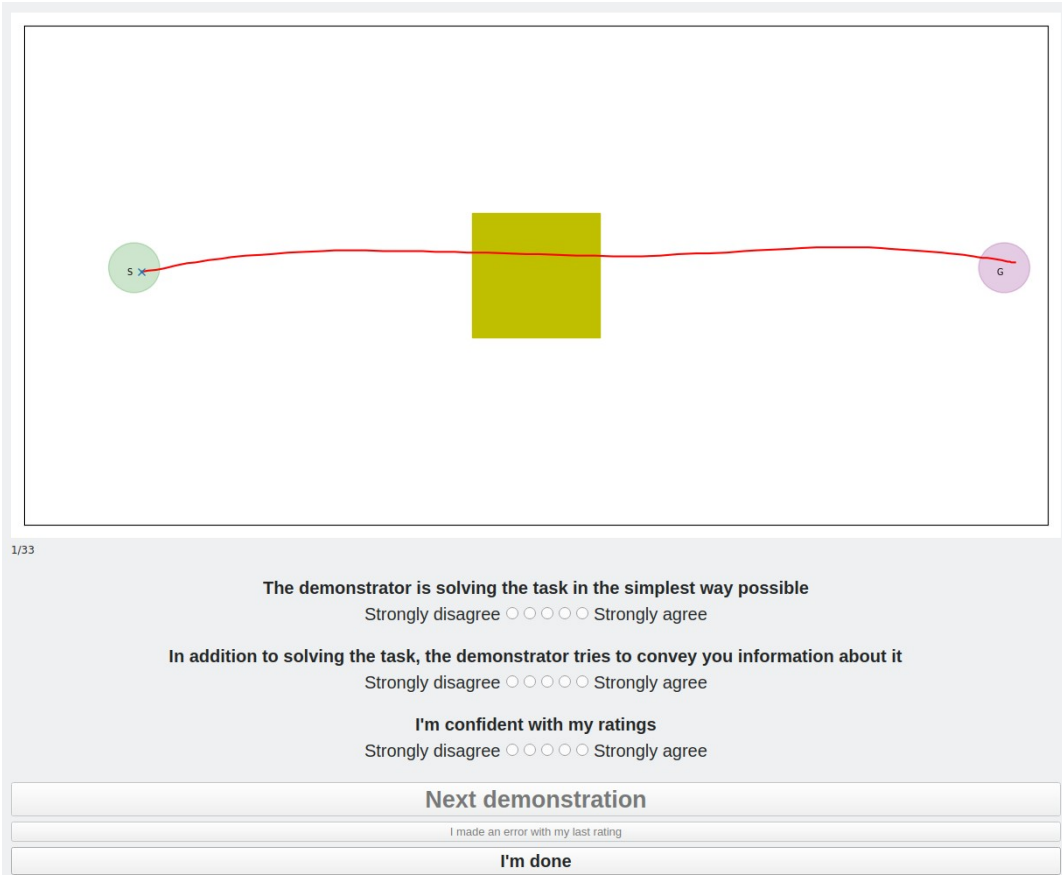


Figure C.3: GUI of experiment 2 (part 3).