UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Explanation-Guided Learning for Human-AI Partnership

Silvia Tulli

Supervisor: *Doctor Ana Maria Severino de Almeida e Paiva*

Co-Supervisors: *Doctor Francisco António Chaves Saraiva de Melo*
Doctor Mohamed Chetouani

Thesis approved in public session to obtain the PhD Degree in

Computer Science and Engineering

Jury Final Classification: Pass with Distinction

2023

TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Explanation-Guided Learning for Human-AI Partnership

Silvia Tulli

Supervisor: *Doctor Ana Maria Severino de Almeida e Paiva*
Co-Supervisors: *Doctor Francisco António Chaves Saraiva de Melo*
Doctor Mohamed Chetouani

Thesis approved in public session to obtain the PhD Degree in

Computer Science and Engineering

Jury Final Classification: Pass with Distinction

Jury

Chairperson:

Doctor Mário Jorge Costa Gaspar da Silva, Instituto Superior Técnico, Universidade de Lisboa

Members of the Committee:

Doctor Tony Belpaeme, Faculty of Engineering and Architecture, Ghent University, Belgium
Doctor Ana Maria Severino de Almeida e Paiva, Instituto Superior Técnico, Universidade de Lisboa
Doctor Maria Inês Camarate de Campos Lynce de Faria, Instituto Superior Técnico, Universidade de Lisboa
Doctor Bradley Hayes, College of Engineering and Applied Science, University of Colorado Boulder, USA

Funding Institution

European Union's Horizon 2020 Research and Innovation Programme Grant Agreement No 765955

2023

For Margherita and Leonardo

Resumo

Sistemas artificiais inteligentes de todos os tipos realizam tarefas complexas e muitas vezes informam os processos de tomada de decisão humana. Consequentemente, necessitam de transmitir informação relevante sobre os processos que levam a determinados resultados assim como a estabelecer um diálogo entre tanto humanos como outros sistemas de IA. De forma a garantir que as máquinas continuem a ser benéficas para os seres humanos requer que estes sistemas continuem a ser capazes de comunicar o seu funcionamento interno, de modo a que outro observador possa inferir o seu raciocínio e intenção(ões). Este processo, conhecido como explicabilidade, é crucial para ajudar a moldar a nossa relação com os sistemas de IA.

Apesar das vantagens das abordagens existentes para implementar a explicabilidade em sistemas de IA e aprender através de interacções mais naturais com humanos e outros agentes, os algoritmos actuais geralmente (1) não são avaliados em cenários de trabalho de equipa e de tomada de decisão humana e (2) requerem frequentemente um grande número de exemplos sobre como resolver uma tarefa. Estes são ambos aspectos cruciais para que os humanos operem ao lado de agentes autónomos, especialmente em cenários interactivos. Para abordar as limitações acima mencionadas, nesta tese realizámos três estudos focados, em primeiro lugar, em compreender o papel das explicações no trabalho de equipa de agentes humanos, em segundo lugar, em explorar a aprendizagem de agentes inteligentes utilizando explicações geradas por máquinas e, em terceiro lugar, em incorporar as explicações humanas na aprendizagem de máquinas.

Começámos por primeiro desenvolver um módulo de transparência num jogo de bens públicos onde um jogador humano pode escolher contribuir para o objectivo da equipa (cooperar) ou agir de forma egoísta para atingir o seu objectivo individual (desertar). Três jogadores, um humano, e dois agentes artificiais jogam juntos. Comparamos os efeitos das estratégias dos agentes (isto é, cooperativas, individualistas, e "olho por olho"), e explicações sobre as suas estratégias nas escolhas de cooperação humana. Encontrámos um efeito de interacção entre a estratégia e a explicabilidade dos agentes na confiança, identificação de grupos e na atribuição de semelhanças humanas, demonstrando que a explicabilidade desempenha um papel fulcral nas colaborações humano-AI.

De seguida, implementamos um jogo para dois jogadores de soma zero chamado Minicomputer Tug of War baseado em linguagem não verbal para introduzir crianças à aritmética mecânica e mental através de notação decimal com regras binárias posicionais. Implementamos o jogo num cenário de aprendizagem de crianças-robô em que o robô explica à criança a suboptimidade das suas acções. Mostramos que as crianças na condição explicável percebem o pós-teste como tendo

menos dificuldade em relação ao pré-teste. Estes resultados mostram que fornecer explicações sobre comportamentos sub-óptimos tem um efeito positivo em cenários de aprendizagem com crianças.

Finalmente, incorporamos explicações na aprendizagem de reforço invertido de máxima probabilidade e desenvolvemos um sistema para gerar explicações contrastivas e avaliá-las contra outros sinais pedagógicos provenientes de um agente especializado (por exemplo, recompensas, demonstrações). Mostramos que o agente que aprende com as explicações em comparação com os sinais de recompensa e demonstração tem um melhor desempenho, indicando que as explicações são uma forma valiosa de transferir sucintamente o conhecimento sobre uma tarefa.

Juntos, estes estudos mostram como os sistemas de IA que exibem uma agência explicável e que são capazes de aprender com as explicações de outros podem afectar positivamente o trabalho de equipa e a tomada de decisões dos agentes humanos, bem como aprender mais eficientemente em comparação com os sistemas não explicáveis. Ao apresentarmos os nossos modelos computacionais em torno destes aspectos, esperamos avançar o nosso conhecimento e compreensão das diferentes facetas da agência explicável em inteligência artificial e permitir uma parceria bem sucedida entre a IA humana e a transferência de conhecimento.

Palavras-chave: IA Explicável, Aprendizagem Máquina Orientada para a Explicação, Aprendizagem Social, Aprendizagem das Demonstrações, Interacção Humano-Robot.

Abstract

Artificial intelligent systems of all kinds undertake complex tasks and often inform human decision making processes. Consequently, they need to convey meaningful information about the processes that lead to a certain outcome and establish a back and forth dialogue with both humans and other AI systems. Ensuring machines remain beneficial to humans requires that these systems are still able to communicate their inner workings in such way that another observer can infer its reasoning and intent/s. This process, known as explainability, is crucial in helping to shape our relationship with AI systems.

Despite the advantages of existing approaches to implement explainability in AI systems and learn through more natural interactions with humans and other agents, current algorithms generally (1) are not evaluated in teamwork and human decision-making scenarios and (2) often require large numbers of examples on how to solve a task. These are both crucial aspects for humans to operate alongside autonomous agents, especially in interactive settings. To address the above-mentioned limitations, in this thesis we conducted three studies centered around first, understanding the role of explanations in human-agent teamwork, second, exploring learning from intelligent agents using machine-generated explanations, and thirdly, incorporating human explanations into machine learning.

We first develop a transparency module in a public goods game where a human player can choose to contribute to the goal of the team (cooperate) or act selfishly in the interest of his or her individual goal (defect). Three players, one human, and two artificial agents play together. We compare the effects of agents' strategies (i.e., cooperative, individualistic, and tit-for-tat), and explanations about their strategies on human cooperative choices. We found an interaction effect between agents' strategy and explainability on trust, group identification, and human-likeness attribution, demonstrating that explainability plays a pivotal role in human-AI collaborations.

We then implement a two-player zero-sum game called *Minicomputer Tug of War* based on a non-verbal language to introduce children to mechanical and mental arithmetic through decimal notation with binary positional rules. We deploy the game into a child-robot learning scenario in which the robot explains to the child the suboptimality of its actions. We show that the children in the explainable condition perceive the post-test as having less difficulty with respect to the pre-test. These results show that providing explanations about suboptimal behaviors has a positive effect in learning scenarios with children.

Finally, we incorporate explanations into maximum likelihood inverse reinforcement learning and develop a framework to generate contrastive explanations and evaluate these against other

teaching signals coming from an expert agent (e.g., rewards, demonstrations). We show that the agent learning from explanations compared to reward and demonstration signals perform better, indicating that explanations are a valuable way to succinctly transfer knowledge about a task.

Together, these studies show how AI systems that exhibit explainable agency and are able to learn from explanations of others can positively affect human-agent teamwork and decision making, as well as learning more efficiently compared to non-explainable systems. In presenting our computational models around these aspects, we hope to advance our knowledge and understanding of different facets of explainable agency in artificial intelligence and enable successful human-AI partnership and knowledge transfer.

Keywords: Explainable AI, Explanation-Guided Machine Learning, Social Learning, Learning from Demonstrations, Human-Robot Interaction.

Acknowledgments

This research can be seen as a meta-reflection of my PhD journey. Leveraging other experiences made all of this possible. It is incredible to realize how many people directly or indirectly had a profound impact on my life and on this research.

First, I must thank my supervisors Mohamed Chetouani, Francisco S. Melo, and Ana Paiva. The combination of their experiences deeply shaped this thesis, broadened my horizons and encouraged me to expand in ways I could never imagine.

Thanks for my examination committee that took the time to discuss my work and accepted it for public defense. Thanks professors Tony Belpaeme and Bradley Hayes for your valuable questions and reports on the draft of this thesis.

The research in this thesis was made possible by the European Union's Horizon 2020 research and innovation program. Thanks for giving me the unique opportunity to be part of an international research and training network. I am grateful for having been exposed to such a flourishing interdisciplinary environment.

Thanks to my fellow graduate students of the Marie Skłodowska-Curie Project for sundry help and fun discussions. Special thanks go to Ramona Merhej with whom I shared the most throughout these years. I wish to thank also Rebecca Stower and Sebastian Wallkötter for their valuable feedback that often extended beyond the realm of the PhD. Thanks to my Talking Robotics teammates, in particular Patrícia Alves Oliveira, and Miguel Vasco, it has been extremely enriching to work with you. Your growth mindset was a constant source of inspiration.

Thanks to David W. Aha, Prashan Madumal, Rosina Weber and Mark T. Keane, for their mentorship and insightful exchanges. It has been a great experience to initiate and organize the workshop on Explainable Agency with you. I would like to extend my sincere thanks to Sarath Sreedharan, and Tathagata Chakraborti, for welcoming me into the Explainable Planning community and being a foundational reference for my work.

My doctoral journey gave me the opportunity to work in two different groups, in two different countries. I had the opportunity to meet brilliant and outstanding people. Thanks to Kim Baraka for your exceptional help and support. Thanks to Shruti Chandra for your life-saving tips. This work would not have been possible without the input from my collaborators. Thanks to the GAIPS team at *Instituto Superior Técnico* for the warm welcome and invaluable support. I am indebted to Marta Couto for facilitating my first study with children. Thanks Maria José Ferreira for your kindness and remote support. Thanks also to the PIRoS group members at *Institut Systèmes Intelligents et de Robotique* for hosting me in such unpredictable times. Thanks to my friend and

colleague Elisa Massi, I'm truly appreciative of your work and your unconditional positive regard.

I would like to thank all my friends including Sarthak, Giovanni, Virginia, Marcella, Sera, and Sooraj. Each of you deserving a dedicated page of this thesis. Thanks Arnaud, you had a great influence on my perspectives. Your support and encouragement was worth more than I can express with words. Finally, I am very grateful to my family. Their unwavering support gave me the solid foundations I needed to stay on my feet and keep moving forward.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Every day we are faced with autonomous intelligent systems that compute optimal routes, perform weather forecasting, and decide what we see on social media. We often rely on their suggestions without questioning them, and are confident enough in the system's ability to maximize its success and achieve our objectives. We grasp the general idea that is behind the functioning of these systems and construct behavior explanations to answer why these systems behave in a certain way [1]. However, whenever the behavior of these systems does not match our expectations, a general idea about how they operate might not be sufficient for us to comfortably delegate certain decisions to them, especially in critical domains like healthcare [2], criminal justice [3] and financial markets [4].

Artificial intelligence systems of all kinds undertake complex tasks, and inform human decision making processes [5]. Consequently, ensuring that machines remain beneficial to humans requires researchers and practitioners in the field to focus not only on achieving high-level learning performance, but also on guaranteeing human understanding, trust, and control of emerging generation of intelligent artificial partners [6, 7, 8]. For AI systems to make a real-world impact in complex domains, these systems must be able to leverage and enhance human expertise. Thus, developing a two-way communication protocol for the Human-AI partnership becomes a foremost priority.

Explanations have profound effects on the probability assigned to causal claims and on how the artifacts' parts or properties are generalized to novel situations. Ergo, explanations represent a valuable way to enable intelligent agents and humans to reciprocally communicate their reasoning and clarify ambiguous situations [9]. A machine-generated explanation should primarily include information about why and how the model under scrutiny produces its predictions/inferences [10]. It should be designed considering the specificity of how the decision-making algorithm operates as well as its embodiment. For example, AI systems that perceive the environment and act autonomously upon that environment, i.e., intelligent autonomous agents, require shifting the interest in explaining how the agent's successive observations affected its decisions [11]. By extension, embodied AI systems that are able to move through the world and affect a physical environment, can explain their intents and goals by using a larger set of communication modalities compared to their non-embodied counterparts [12].

Generally, an explanation should help in constructing a conceptual framework to interpret the behavior of the AI systems, debug [13] and eventually allow the human to select informative examples to instruct AI systems to solve tasks that occur unexpectedly [14]. Additionally, whenever the AI systems embed information that another artificial or human agent might be unaware of, explanations might suggest different strategies to solve a task and lead to a deeper

understanding of a problem.

Just as humans use explanations to teach each other, intelligent agents could potentially do the same. Explanations would therefore not only be a means of designing transparent systems, but also a more efficient way to transfer knowledge from humans to agents, and vice-versa. Through learning from explanations and generating explanations for learning, agents could make other artificial and human agents learn better and faster than they would by trying actions on their own. Despite the advantages of existing approaches to implement explainability in AI systems and learn through more natural interactions with humans and other agents, current algorithms generally (1) are not evaluated in teamwork and human decision-making scenarios and (2) often require large numbers of examples on how to solve a task. These are both crucial aspects for humans to operate alongside autonomous agents, especially in interactive settings. To address the above-mentioned limitations, in this thesis we conducted three studies centered around first, understanding the role of explanations in human-agent teamwork, second, exploring learning from intelligent agents using machine-generated explanations, and thirdly, incorporating human explanations into machine learning. In presenting our computational models around these aspects, we hope to advance our knowledge and understanding of explainability and of the central role it plays in both human-AI partnership and knowledge transfer.

1.1 Thesis Overview

1.1 Research Questions

Explanations in Collaborative Settings

Despite the interest in explainability, less attention has been placed on the effects of explanations in collaborative situations, where revealing the strategies of others may affect the choices of each member of a team. This thesis explores this situation in human-agent teams. First, we develop a transparency module to explain the strategies of three different artificial players in a public goods game. Then, we observe human cooperative choices depending on the artificial players' strategies and explanations. We specifically address the following research question:

- **RQ1** Does explaining the strategies of agents in human-agent teams foster more collaborative behaviors in the human?

Explanations in Learning Scenarios

Although feedback and demonstrations have been largely investigated in machine learning scenarios [15, 16, 17, 18, 19, 20, 21], the design and evaluation of agents' explanations to foster knowledge transfer from humans to agents and vice-versa has hardly been explored. This could be explained by two different classes of problems. First, the fact that learning from intelligent autonomous agents implies having agents that are capable of providing explanations about their inner workings specifically for human learners, i.e., ability to represent task in a more succinct form that makes sense for the human. Second, it implies having intelligent autonomous agents that can reason human explanations, i.e., ability to make sense of a concise representation of a task that would also makes sense for the agent, which is already challenging itself, especially in complex domains [22, 23].

The first class gathers problems related to the representation of the agent's explanations. On this regard, our work is inscribed within methods that aim at evaluating the utility of alternative plans as an indicator of the degree of sub-optimality of the performed action. The second class is related to the agent's reasoning. In regard to this, our work mainly focuses on integrating explanations into maximum likelihood inverse reinforcement learning as a means to provide information about the goodness of possible states and actions.

In particular, we study the following research questions:

- **RQ2** How can intelligent autonomous agents provide explanations about their behaviors to enhance human understanding of a new task?
- **RQ3** How can intelligent autonomous agents learn from another agent's explanations?
- **RQ4** Does explanation compared to demonstration and reward signals lead to better learning?

1.1 Contributions

This thesis holds four major contributions (Figure 1.1):

- **A comprehensive taxonomy of both the desiderata and methods in explainability research with a specific focus on sequential decision-making agents and embodied agents**. There is little consensus on what explainability is and how to evaluate it for benchmarking. We reviewed existing definitions of explainability and stated our. In particular, we explored the topic of explainability in embodied agents claiming that embodied agents have access to differing social cues compared to their non-embodied counterparts. Chapter

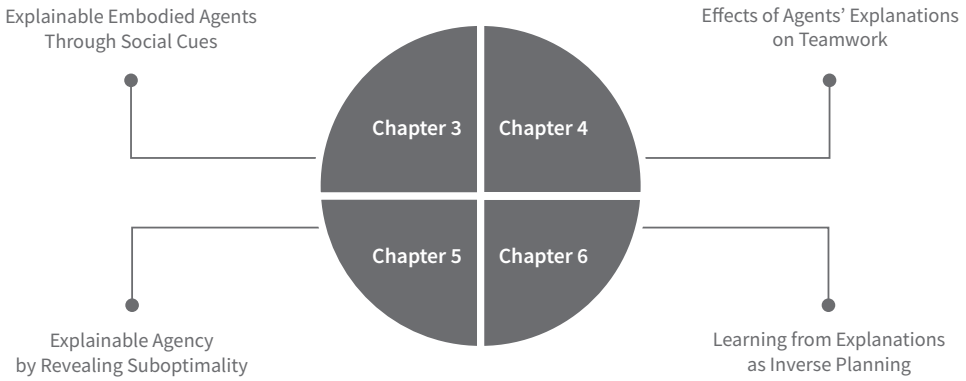

Figure 1.1: Outline of the thesis.

3 details our contribution towards the understanding of explainability by showing how explainability is implemented and how it exploits different social cues.

- **An evaluation of the effects of agent's explanations on teamwork**. When autonomous systems move from being tools to being teammates, an expansion of the interaction model is needed to support the paradigms of teamwork. Explainability facilitates the understanding of the responsibilities that different group members might take in collaborative tasks. For this reason, combining the investigation of the behavioral model of the players in relation to the different strategies of the team members and the explainability of the decision-making process of the artificial players turns out to be useful for the design of systems that aim to facilitate and foster collaboration. Chapter 4 addresses **RQ1** and presents a user study that investigates the effect of the explainability and strategy of virtual agents on human collaborative choices.

- **An implementation and evaluation of a computational model for generating explanations of an agent's plan**. Revealing the internal workings of an agent solving a task can help another agent better understand the task. How to reveal such workings, e.g., via explanation generation, remains a significant challenge. This gets even more complex when these explanations are targeted towards children. We propose a search-based approach to generate contrastive explanations using optimal and sub-optimal plans and implement it in a child-robot scenario, answering **RQ2**. Chapter 5 reports results around our explanation generation system that was successfully deployed among seven-year-old children.

- **An implementation and evaluation of a computational model for learning from other agents explanations**. Evidence from inferential social learning suggests that humans draw flexible inferences by building structural causal models of how other agents' hidden states cause their actions. We plan to extend the learning capabilities of agents by leveraging the knowledge of other agents. We discuss methods to learn from explanations in inverse reinforcement learning. We argue that explainability methods, in particular methods that use counterfactuals, might help increase sample efficiency. Chapter 6 address questions **RQ3** and **RQ4** by discussing empirical results around humans' preferences for different kinds of explanations, and different approaches to learn from explanations and other teaching signals.

1.1 Roadmap

The remainder of this thesis is organized as follows. **Chapter 2** concerns an overview of the related literature including theories of explanation in philosophy, cognitive and social sciences, definitions and desiderata in explainability research, imitation learning approaches, early work on explanation-based learning and recent research on learning rewards from explanations. We give a particular emphasis on explainability methods in collaborative environments, embodied agents, and sequential decision making agents. **Chapter 3** includes a literature review on the specific case of explainable embodied agents. **Chapter 4** reports our first experimental study aimed at investigating if agent's explanations promote humans collaborative choices in a game scenario. **Chapter 5** discusses the topic of explainable agency for human learning, reporting a user study in which we generate the robot's explanations by comparing optimal and sub-optimal actions.

Chapter 6 presents the final experimental studies in which we measure humans' preferences for different kinds of explanations, and evaluate a computational model for learning from explanations using maximum likelihood inverse reinforcement learning. **Chapter 7** positions our contributions and discusses the current and potential future impacts the work can have in the explainable AI landscape. We conclude with a reflection for exciting avenues extending the research presented in this thesis toward the development of explanation-guided learning and explainable learning agents for human-AI partnership.

Chapter 2

Related Work

This chapter provides an overview of topics at the intersection between explanations, explainability and machine learning. First, it defines the structure and the function of explanations in philosophy, psychology and cognitive science. Second, it summarizes the desiderata in explainability research with a special focus on explainability methods for intelligent agents and robots. The focus on agents and robots is motivated by the fact that the sequential nature and/or embodiment of these agents poses challenges that other machine learning techniques like classification do not face. Further, the chapter describes existing methods for learning from other intelligent agents, encompassing imitation learning methods and including early work on explanation-based learning. Since explainable agency is an attribute given by the observer and inherently motivated by the human quest of understanding the agents' behavior, further research exploring human-agent join activities and human-agent teams is discussed. In each section, an overview of existing gaps, and how the thesis aims to contribute to the current state of the art is provided.

2.1 Structure and Function of Explanations

Humans always wonder why a situation unfolds in the way it does, why objects have specific properties, or why someone takes certain decisions. While seeking for explanations, humans pursue the necessary understanding of a problem's solution that enables them to then generalise these solutions to novel situations. While there are vast and valuable bodies of research from philosophy, psychology and cognitive science exploring the human need for explanations, the definition of explanations turns out to be complex and multifaceted. An *explanation* has been defined as: (1) an answer to why or how-questions [24]; (2) hypotheses about possible causes behind the object of the explanation [25, 26]. The *act of explaining* can be thought as a means to transfer knowledge between an *explainer*, i.e., someone who is in possession of explanatory information, and an *explainee*, i.e., someone or a group of people who is thought not to possess it yet [27, 28]. This process has been identified as the *social process* of the explanation [29]. In a continuous interaction between the explainer and her counterpart, the main goal of the explainer is to provide enough information to the explainee so that they can understand the causes of some fact or event. This process contemplates the active role of the explainee, who can ask for explanations by querying the explainer.

In addition to the *social process*, explanation has also been described as a *cognitive process* and a *product* [9, 30]. The *cognitive process* concerns abductive inference, a form of logical inference that starting from the observation or set of observations seeks for the simplest and more likely

conclusion , i.e., explanatory hypotheses [31, 32]. Throughout this inference process, one selects the potential causes to build an explanation for an *explanandum*. The potential causes for an *explanandum* are called *explanans*. The *explanans* constitute the *product* of an explanation, as first explained by Hempel and Oppenheim [33] (Figure 2.1).

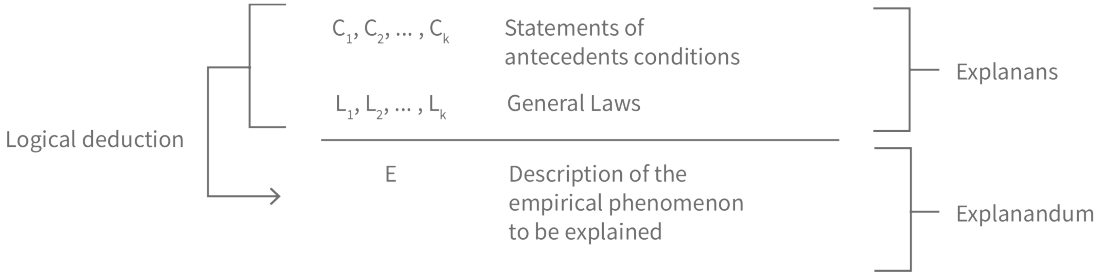

Figure 2.1: The basic pattern of scientific explanation introduced by Hempel and Oppenheim [33]

Lombrozo [9, 34] identifies evidence from cognitive psychology and cognitive development concerning the structure and function of explanations, with a focus on the role of explanations in learning and inference. Their research indicates that explanations can have profound effects on the probability assigned to causal claims and on how artifacts parts or properties are generalized. Logical or causal constraints constitute the pillars of an explanation. Influential accounts state that what has been explained, i.e., the *explanandum*, follows from natural laws or empirical conditions and it is an instance of general patterns or regularities [33, 35]. Prior knowledge of general patterns aid the selection of which causes are judged probable and relevant. These causes are referred to as *explanatory*. The identification of relevant properties of an *explanandum* provide a principled basis for generalization. The ensemble of general patterns define and create prior beliefs which are used to exclude inconsistent possibilities and serve as a source to constraint reasoning.

Explanations have been closely linked to reasoning, and understanding. Aristotle stated that *"we think we have knowledge of a thing only when we have grasped its cause*". He conceives the causal investigation as a request for explanations, and identified four fundamental types of causes, or *modes of explanation*: material, formal, efficient, and final. The material cause is the substance of something or what it is made of , i.e., the table is made of wood, therefore the wood is the material cause of the table. The formal cause is the structure or properties of something , i.e., something that has a flat top and one or more legs might be a table due to its form. The efficient cause represents the external agent responsible for the change in the material to obtain its desired form , i.e., the carpenters is the efficient cause of the table because they gather the materials and fashions them into the table form. The final or functional cause describes the purpose of

something or *"that for the sake of which*" , i.e., the table is used as a surface for working at or eating from, thus working and eating are the functional cause of the table.

Several scientists have advocated that there exist systematic explanatory preferences. Evidence from cognitive science supports the hypothesis that the psychological function of explanation is to highlight information likely to subserve future prediction and intervention [36]. Research on whether simple causal explanations are preferred to more probable ones suggests that *simplicity*[1] plays a privileged role in assigning prior probabilities, i.e., simpler explanations are preferred and judged more likely. Moreover, evaluating explanations may serve as a mechanism for generating estimates of *subjective probability*[2] [39].

These findings emphasize that generating and evaluating explanations have an instrumental role in constructing prior beliefs about a task, especially in knowledge-rich domains. Consequently, the study of explanations opens a unique window onto foundational aspects of cognition ranging from conceptual representation to learning and inference [34]. Human studies on self-explanation, e.g., elaborating on a topic by explaining to oneself, affirm that self-explanation scaffolds causal learning and problem solving. The self-reflection process bringing to the formulation of an explanation is further strengthened in the social learning context. When teaching others, humans tend to spend more time reviewing their own knowledge. The mere belief of a social interaction lead to assigning a greater value to an explanation helping to cement the learning of new associations [40]. This sense of responsibility towards others that motivates learning has been called *protégé effect* or *learning by teaching* and it has been proven to provide an environment in which knowledge can be improved through revision [41, 42].

It follows that in social contexts, as in teacher-student interactions, an explanation goes beyond the explainer prior beliefs about a task. The explainer makes considerations about its audience to select information and communication modalities that can resonate with it. Ergo, the success of an explanation depends on several critical factors belonging to the *explainee* or *audience*; the audience's assumptions about a task, their previous knowledge, and interests. The social psychology perspective offers yet another lens for investigating explanations [28, 1].

The primary focus of this thesis will be on developing novel explainable models for intelligent agents grounded on insights from psychological and philosophical analyses on how humans formulate explanations and reason about causal claims. In addition, another aim of this thesis is to investigate how findings about explanations from social sciences generalize to intelligent artificial agents, and explore how explanations contribute to transfer knowledge and efficient

[1] Following Newton's definition [37], *simplicity* is the measure of the number of causes invoked in an explanation, while *probability* quantifies the more probable among these.

[2] A subjective probability is anyone's opinion of what the probability is for an event [38]

teamwork in human-AI partnership. Finally, this thesis implements methods to endow agents with the ability to reason upon others' explanations to allow them to differently interpret future prediction and intervention, thus learn more efficiently as humans do.

2.2 Explainable AI

Solutions to generate explanations about the inner workings of AI systems has been widely studied under the umbrella of Explainable AI. Much of recent progress in Explainable AI (XAI) concentrates on developing interpretable machine learning models. Lipton [43] suggests that *interpretability* is a prerequisite for five important aspects: trusting a model, inferring causal relationships, transferring learned skills to unfamiliar situations, providing actionable information, and supporting fair and ethical decision-making. Their research debates the definition of trust, and affirms that *how often a model is right* but also *for which examples it is right* may be more relevant that just providing information about the confidence or accuracy of a model. Furthermore, a model is *trustworthy* when there is no expected cost of relinquishing control, i.e., the model is as accurate as the human. Although the associations learned by supervised learning models do not guarantee potential causal relationships, developing model interpretability has been identified as a way to provide clues about genuine and spurious causes [3] responsible for both associated variables. This aspect has been largely investigated in probabilistic graphical models for learning causality from observational data [45] such as *Bayesian Network*.

Interpretable methods developed for explaining the outputs of supervised and unsupervised machine learning models have been also employed in sequential decision making models, e.g., SHAPLEY values [46] and saliency maps [47]. However, the sequential nature of these models make them different from other machine learning systems (e.g., classification) [48], posing new challenges for explainable AI research. Differently, sequential decision making systems map perceptual inputs from the environment to a sequence of actions. It follows that explaining a one-shot decision, e.g. how much an input affected a certain prediction, might not be sufficient for making the decision-making process of these models intelligible for the human. To be explainable, sequential decision making models, e.g., reinforcement learning agents, have to explain their actions accounting for the dependency that their actions have with previous actions, rewards and environmental conditions.

[3] spurious correlation occurs when two factors appear casually related to one another but are not [44].

2.2 Explainable Agency

Explainability in sequential decision making agents has been often referred to as “explainable agency". Whenever pursuing human-specified objectives, explainable agents should be able to summarize their activities and answer questions about the reasons for their interdependent decisions [49]. Given a set of objectives and the necessary background knowledge that is relevant to these objectives, to be explainable an agent should (1) produce records of decisions made during its reasoning, (2) summarize its behavior in human accessible terms and (3) provide answers to questions about specific choices and the reasons for them [50]. Producing records of decisions made during planning should include stating the alternatives the agent considered, giving its reasons for selecting them over alternatives, and describing its expectations for each option [51]. The information provided by the agent needs to be expressed at different levels of abstraction as appropriate and clarify how the performed actions relate to inferences made by the agent. As previously stated, explanations should be given especially in situations in which actual events diverged from expectations and the agent had to adapt in response. To ensure intelligibility, the information should be presented in terms of beliefs, goals and activities that people find to be familiar. The three main architectural components an explainable agent must include are: (1) a representation of content that support explanation, which might require symbolic structures; (2) an episodic memory to collect records of relevant information; and (3) the ability to access and extract content from these records [52].

Several attempts have been made to develop explainable agency in intelligent agents. We identified two distinct approaches: causal, and non-causal approaches. Causal approaches focus on differentiating properties of two competing hypotheses [53, 54, 55]. In contrast, policy summarization approaches provide a description of the agent’s global behavior without requiring a specific contrast case [56, 57]. Another strand of research focuses on the interpretation of human queries by either mapping inputs to query or instruction templates [58, 59, 60], or by using an encoder-decoder network to learn associations between natural language behavior descriptions and state-action information [61].

Causal Approaches

Causal approaches draw from the notion that humans use causal, not statistical, logic. Furthermore, causal reasoning, which deals with *what-if*, is essential for intelligent agents to communicate with humans about their policies and future intentions [62]. Counterfactuals have been seen as a fundamental part of causality since the beginning of this field [63]. Providing

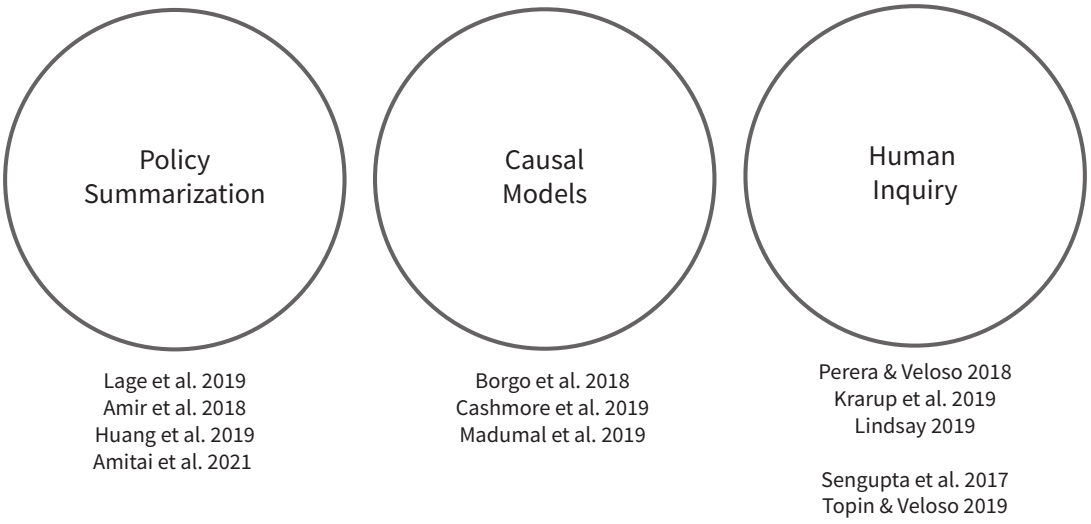

Figure 2.2: Overview of Explainability Methods for Intelligent Autonomous Agents

explanations in the form of counterfactuals (e.g., “If A had not occurred, C would not have occurred”) or contrastive explanations (e.g., explanations which answer to “Why P rather than Q?") has received a lot of attention [64, 48, 65, 66]. By reasoning over counterfactuals, i.e., other possible configurations of the world, humans compare alternative representations that eventually affect future intentions and decisions [67]. As a consequence, endowing agents with the ability to reason and plan over alternative configurations of the world where more rewarding optimal policies may be possible has been studied both to improve the agent planning under uncertainty [68] and to generate explanations about alternative plans that make sense for the human [69, 70, 71].

Borgo et al. [69] investigated this aspect by developing a methodology for comparing the cost of the robot’s plans and allowing the user to investigate alternative actions within them. Cashmore et al. [70] showed how to incorporate human suggested action by adding, changing or removing actions from the planner’s original plan, and comparing the cost of both plans. In extension, the work of Tsirtsis et al. [71] proposes to go beyond counterfactual explanations for one-step decision making processes [72, 73] by introducing a method to find counterfactual explanations in situations in which multiple, dependent actions are taken sequentially over time. In their formulation a counterfactual explanation specifies an alternative sequence of actions differing in at most k actions from the observed sequence that could have led the observed process realization to a better outcome. By using synthetic and real data from cognitive behavioral therapy, they have shown that their polynomial time algorithm to find optimal counterfactual explanations can

provide valuable insights to enhance sequential decision making under uncertainty.

Juozapaitis et al. [74] study reward decomposition[4] for the purpose of explanation and focus on pairwise action explanations where the goal is to explain why one action is preferred to another in a particular state. They focus on explanations for RL agents that learn q-functions where the q-function gives the expected infinite-horizon γ-discounted cumulative reward[5] of taking action a in state s and following a policy π. A q-function is decomposed in q-vectors consisting of the q-values for each state-action pair. The authors define *reward difference* explanations (RDX) as the difference of the decomposed q-vectors $\Delta\left(s, a_1, a_2\right) = \vec{Q}\left(s, a_1\right) - \vec{Q}\left(s, a_2\right)$. Each component of the RDX is a positive or negative reason $\Delta_c\left(s, a_1, a_2\right)$ for the preference depending on whether a_1 has an advantage (disadvantage) over a_2 with respect to reward type c. The *minimal sufficient explanation* (MDX) is a more compact version of RDX and comprises of a small set of the most important reasons about why an action is preferred to another.

In extension, Madumal et al. [11] formalize an *action influence model* to learn the quantitative influences that actions have on variables of interest of a task. Their computational evaluations are accompanied with a user study to measure the participants' performance in task prediction, explanation satisfaction, and trust.

Nonetheless, explaining why a given solution is better than an alternative by comparing their costs or utilities may not be sufficient. To overcome this limitation, other approaches have focused on providing an answer to different but equally important aspects of the decision-making process. Examples of such approaches include detailing how to reach an optimal state, describing the consequences of possible errors, and interpreting reinforcement learning policies by looking at the results of the training and final solutions [75].

Motivated by the previously mentioned casual approaches, we generate explanations to answer counterfactual questions, e.g., what caused the observation(s) to occur? Why is this state/action better with respect to another?

In contrast to previously mentioned approaches, in Chapter 6 we uses maximum likelihood estimation to infer the parameters of an expert's reward function and integrate these estimates into inverse reinforcement learning. By using inverse planning we are able to provide the agent with the ability to reason upon the generated explanations. Furthermore, by using a probabilistic approach we can fairly compare learning from explanations against learning from other teaching signals (i.e., rewards, demonstrations) and provide evidence on the effect of explanations in

[4]Reward decomposition consists in decomposing a reward function by specifying a set of reward components/types

[5]infinite-horizon means that the agent considers infinite steps into the future when receiving the reward, γ refers to the discount factor determining how much the RL agents cares about rewards in the distant future relative to those in the immediate future. If γ=0, the agent will be completely myopic and only learn about actions that produce an immediate reward.

agent's learning performance.

Policy Summarization

Policy summarization can be considered as a way to exhibit explainable agency without causal implications. It helps highlight the agent's capabilities and weaknesses (1) by demonstrating actions taken by the agent in different states [56], (2) by extracting state-action pairs useful for recovering the agent's reward function [76, 57], or (3) by comparing agent policies [77]. Important states, state-action pairs or policies are selected based on agent q-values and state similarity, quality of the reconstructed policy, or agents' disagreement states, respectively. Similarly with policy summarization approaches, in Chapter 5 we discuss approaches to generate explanations that highlight the weaknesses of the agent by comparing agent policies. On the contrary, we measure the effects of these explanations on children understanding of a task.

Human Inquiry

Other works enable humans to inquire about the possible causes of the planning results. Early work on explaining a system's action already began to investigate how expert systems [78, 79, 80], or semantic nets [81, 82], which use classical search methods, could integrate explanation in interactive scenarios to allow both, inquiry about the decision in the form of specific questions, and rule-extraction based on previous decision criteria [80]. Connecting to this history, the majority of recent research which include human inquires retrieves information about previous states and actions and uses these to answer questions about the robot's plan, e.g., [83, 84, 59, 85]. In this context, parallel research have been extending this approach to queries concerning future actions as well [86, 87]. In this context, RADAR is a particularly interesting approach, because it focuses on proactively aiding and alerting the humans in the loop with their decisions rather than generate a static plan that may not work in the dynamic worlds that the plan has to execute in [86].

2.2 Explainable Embodied Agents

In the above mentioned explainability methods, explanations are often provided in the form of template-based sentences. However, when the AI systems are in possess of an embodiment, the spectrum of modalities that can be used to provide explanation is much broader compared to their non-embodied counterparts.

Research in explainable embodied agents refer to transparency, explainability, expressivity,

understandability, predictability and communicability [12] to define the ability of a system to convey its inner workings.

Some authors refer to transparency as the process or the capability of revealing hidden or unclear information about the agent. Akash et al. [88][89] and Ososky et al. [90] cited the definition of Chen et al. [91], which named transparency the descriptive quality of an interface. The descriptiveness of the interface affects the human operator on three levels of awareness about the agent. These levels, leveraged by Endsley's model of situation awareness [92] (Figure 2.3), concern respectively what the human knows about the agent's (1) current status, actions and plan, (2) reasoning process, and (3) projection, predictions and uncertainty. Ososky et al. [90] further refer to the dictionary definition of *transparency*[6]; thus, they chose the property of being *able to be seen through* or *easy to notice or understand* as their definition. Chao et al. [93] gave a similar definition in the context of robot active learning referring to transparency as revealing to the teacher unknown information about the robot.

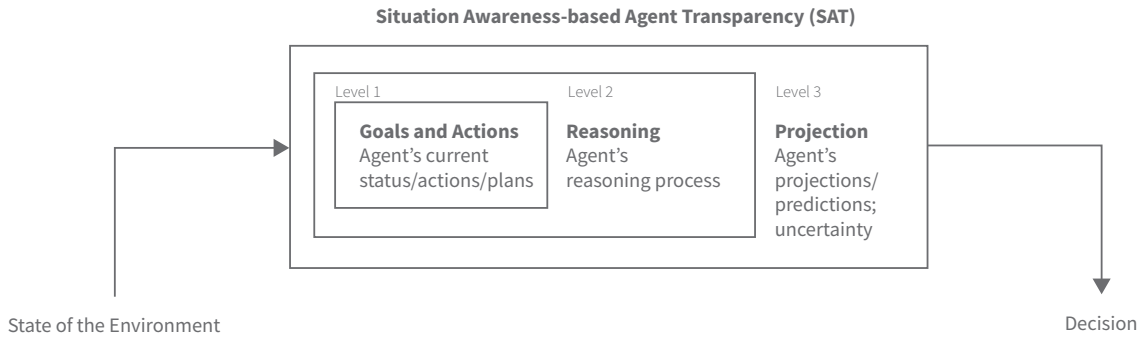

Figure 2.3: Chen's model of Situation Awareness-based Agent Transparency [91]

In an extension, Floyd and Aha [94] and Hayes and Shah [58] refer to both explainability and transparency. Hayes and Shah [58] describe explainability as the embodied agent's ability to synthesize policy descriptions and respond to human collaborators.

Starting from the research of Kulkarni et al. [95] and Zhang and Liu [96], Chakraborti et al. [97] and [98] formulated their idea of explainability as a model reconciliation problem. They use explanations to move the human model of the robot to be in conformance with the robot model (Figure 2.4). Gong and Zhang [99] differentiate their work by shifting the interest in signalling with behavior explanations the robot intentions before actions occur. Along similar lines, Tabrez et al. [100] defined explainability as a policy update given by the robot to the human to reduce the likelihood of costly or dangerous failures during joint task execution.

[6]https://www.merriam-webster.com

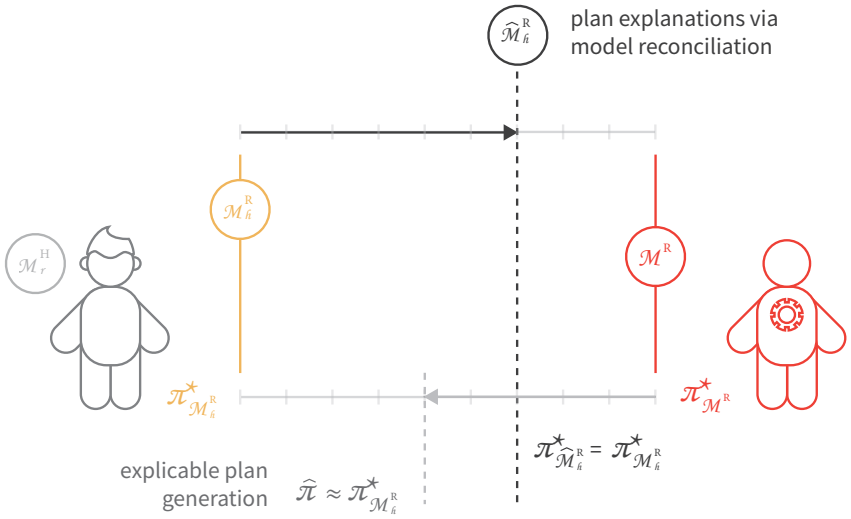

Figure 2.4: Plan Explanations via Model Reconciliation [98, 97]

Baraka and Veloso [101] employed the term *expression* for *externalizing hidden information of an agent*. The work of Kwon et al. [102] extended this notion of expressivity by targeting the communication of robot incapability. To do so, the robot should reveal the intentions and the cause of its failures. The same concept is referred to using the word communicability; e.g., Huang et al. [103] refer to *communicate* for describing the robot capability of expressing its objectives and the robots capability to enable end-users to correctly anticipate its behaviour.

Similarly, Schaefer et al. [104] investigated the *understandability* of the embodied agent's intentions for effective collaborations with humans. Following this idea, Grigore et al. [105] referred to *predictability*, building upon hidden state representation.

Other studies in the literature do not refer to a specific capability of their system in the title but highlighted the application scenarios (e.g., autonomous driving [106], human robot teams in the army [107], interactive robot learning [108]) or mentioned *behavioural dynamics* [109] and *human-machine confidence* [110] to refer to similar concepts).

Although there exists large diversity and inconsistency in the language, there seem to be commonalities in what the authors identify as transparency, expressivity, understandability, predictability and communicability, and explainability. We noticed that all the given definitions share the following aspects: (1) they all refer to an agent's capability or system's module, (2) they all specify that what should be explained/signalled are the internal workings of the agent (e.g. agent's intent, policy, plan, future plans), and (3) they all consider the human as a target of the explanations.

While investigating the definitions, we noticed that the motivation of the experiment plays a

key role in the choice of a specific definition. In particular, we identified the following reasons for investigating explainability:

- **Interactive machine/robot learning** investigates the need of explainability in the context of machine/robot learning. The main idea is that revealing the agent's internal states allows the human teacher to provide more informative examples [108, 93, 111, 112].
- **Trust** states that adding explainability increases human trust and system reliability. This motivation empathizes the importance of communicating the agent's uncertainty, incapability, and existence of internally conflicting goals [113, 114, 102, 104].
- **Teamwork** underlines the value of explainability collaboration scenarios to build shared mental models and predict the agent's behaviour [76, 110, 97, 58, 115].
- **Ethical decision-making** suggests that communicating the agent's decision-making processes and capabilities, paired with situational awareness, increases a user's ability to make good decisions [116, 102, 88].

Guidance and dialogue with a human tutor are aspects that are important for interactive machine/robot learning [108, 93]. Providing information about the level of uncertainty and expressing robot incapability are core concepts of explainability that enhance human trust [113, 114], "express robot incapability" [102]). The ability to anticipate an agent's behaviour and establish a two-way collaborative dialogue by identifying relevant differences between the humans' and the robots' model are shared elements of the definitions around teamwork [76, 110, 117, 58]. Authors that refer to ethical decision-making identify the communication of intentions and context-dependent recommendations as crucial information [116, 89].

2.3 Learning from Others' Experience

When the intelligent agent's explanations are aimed at transferring knowledge about a task, they can be seen as a method to optimize learning.

2.3 Imitation Learning

Learning to perform a task based solely on one's own experience can be unfeasible and costly. Humans address this problem by observing others' explicitly demonstrating or simply showing previously explored solutions, and quickly inferring the appropriate actions to take based on their observations. This social learning mechanisms, known as *imitation learning* (IL), speed

up the acquisition of new skills by reducing the search for a possible solution, by either starting the search from the observed good solution, i.e., *local optima*, or conversely, by excluding bad solutions from the search space [118].

Accordingly, enabling machines to learn a desired behavior by imitating an expert's behavior, being her an expert about how to perform a specific task but not necessary about how to program such a skill into a robot, can be a powerful tool [119]. IL finds its application in many sequential decision problems including continuous and discrete optimization problems, and it is particularly useful in control problems where writing down the reward function that specify how different desiderata should be traded off is challenging [120], e.g., driving a car [121], modeling human intents to navigate in a crowd environment [122], synchronizing lips of a cartoon characters [123].

Imagine having to define the reward function for an aerobatic helicopter flight [124]. The reward function would consist of many features describing relevant aspects for controlling the helicopter, e.g., desired velocity, centripetal acceleration, pitch and so on, that would be difficult to detail using rewards. By recording a pilot's flight and using imitation learning, we can obtain the reward weights that result in policies that bring us closer to the expert, and drastically reduce the online computational cost of the learning problem.

In IL, the learning problem is formulated as a supervised learning problem in which a policy can be obtained by solving a simple regression problem. Differently from classic supervised learning, instead of predicting a single independent and identically distributed random variable (i.i.d.) at the time, IL aims at predicting a sequence of examples. Starting from a set of expert's demonstrations $\mathcal{D}$, usually given as trajectories, the IL algorithm select an appropriate policy representation π_θ, e.g., linear regression, and define an imitation loss function $\mathcal{L}$, e.g., squared error, that represent the mismatch between the demonstrated behaviors and the learner's policy. The policy parameters θ are then optimized for a policy within the policy class, the loss function and the set of demonstrations.

Learning from Demonstrations

Learning from Demonstrations (LfD) is an approach to policy learning. Within LfD, a policy is learned from *examples*. [18, 125]. In contrast to reinforcement learning (RL) approaches that gather data from direct *exploration*, in LfD the data are extrapolated by the teacher's demonstration, thus limited to the states encountered, and the corresponding actions performed, in the demonstration.

One limitation of traditional LfD methods is that they imperfectly capture more abstract, but

equally important information about a skill.

Work on imitation learning can be divided into directly replicating desired behaviors, i.e., *behavioral cloning*, *interactive direct policy learning*, and learning the intents of the desired behaviors, i.e., *inverse optimal control* [126] or *inverse reinforcement learning* [127]. A third paradigm involves the intervention of an interactive demonstrator that can be query, i.e., direct policy learning.

Behavioral Cloning

Behavioral cloning (BC) is the simplest form of imitation learning and consists in mimicking an expert behavior by learning a direct mapping between states and actions without recovering the reward function [128, 129]. Defining P^* as the distribution of states visited by the expert $\mathrm{P}^* = \mathrm{P}\left(\mathrm{s} \mid \pi^*\right)$, e.g., the game screens collected when the expert plays the game, BC encodes the demonstrations as state-actions pairs, and then trains using supervised learning to optimize its objective function, e.g., choose a sequence of actions that minimize the imitation loss. BC offers a simple and efficient alternative to standard supervised learning. However, this approach is particularly sensitive to mismatches in the distribution between training and test samples due to the assumption that state-actions pairs are i.i.d. examples. This assumption breaks in contexts where outcomes are partly random and partly under the control of a decision maker, e.g., MDPs, leading to new states that are assumed to be i.i.d. from P^* even though they are not, thus resulting in an undefined behavior.

Therefore, BC is suitable for (1) learning reactive behaviors, e.g., spam filter, (2) situations in which 1-step deviations do not lead to catastrophic error, and (3) problems in which the expert trajectories cover most of the state space.

Inverse Reinforcement Learning

Inverse Reinforcement Learning is motivated by situations where the knowledge of the rewards is a goal by itself, as in *preference elicitation*, and by the task of *apprenticeship learning*, i.e., learning policies from an expert [130]. Given an MDP without a reward function MDP\R and a set of (near) optimal demonstrations from an observed behavior analyzed as state-action pairs $\mathcal{D} = \{\tau_1, \ldots, \tau_m\} = \left\{\left(s_0^i, a_0^i, s_1^i, a_1^i \ldots\right)\right\} \sim \pi^*$, and the dynamics of the system, the inverse reinforcement learning problem (IRL) [131] is to find a reward function r^* that can explain that behavior and then run reinforcement learning on the inferred reward function to learn a policy

π^* [132]:

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}} \mathbb{E}_\pi \left[r^*(s,a) \right] \tag{2.1}$$

In practise, the agent goal is to infer what task the expert's policy is trying to illustrate (Figure 2.5. One way to compute r^* is by somehow inverting the Bellman equation. By taking a probabilistic approach to the IRL problem, we assume that the demonstrations are a dataset generated by a suboptimal demonstrator, i.e., sometimes the expert can make mistakes. The learner has to recover the weights ω^* associated with linear combination of features describing the $r_\omega(s,a) = \omega^T \phi(s,a)$. After estimating the reward function, RL uses that estimate to compute a policy. The learned policy is then compared with expert's policy and iteratively improved until it converges. However, the expert's demonstrations might not include all the situations that the learning agent might encounter. Therefore, uncovering the optimal reward function might be ambiguous.

IRL is considered an ill-defined problem for two main ambiguities: (1) one reward function can be representative of multiple optimal policies, (2) one policy can encapsulate multiple reward functions.

To solve the ambiguity of finding which of the possible reward functions is the one representing the expert's policy, the work of Abbeel and Ng [133] relaxes the IRL objectives and sets the goal to finding a reward function so that the performance of the optimal policy with respect to that reward function is not much worse than the expert's performance. This approach assumes that (1) the dynamics of the system $\{P_a\}$ is known, (2) we can access to some oracle that given the reward function can solve the MDP, thus implying that the state space is small enough to do that efficiently, and (3) the reward function is linear for the known features of states and actions.

By analyzing the IRL in the terms of finding a policy that matches the expert's features, we are reducing the IRL into a feature matching problem. Since the expert's features cannot be estimated from a limited number of expert's demonstrations, we can further relax our goal so that the difference between the expert's features estimate and the expert's features differ by no more than some ϵ, hence guarantee that the performance of the derived policy is close enough to the expert's policy $\phi(\pi) \approx \phi(\pi^*)$. This formalization can be viewed as a form of regularization.

An example of a regularization scheme is *maximum entropy*. Considering that there exists multiple reward function that correspond to the same policy, it is possible to obtain a stochastic-mixture of policies and still satisfy the features' matching requirement. Consequently, the ambiguity about which of these policies is the expert's policy remains. Maximum entropy provides a principled way to resolve this ambiguity. Starting from a set of distributions over trajectories $P(\tau)$ of a policy π, we can rewrite the feature expectation matching requirement as

the sum of all possible trajectories induced by π. The sum should match the expert's features and should sum up to 1 to be a proper distribution. Following the maximum entropy principle to choose the distribution that better explain the expert's features, we select the one with the largest entropy [134]. Given that the expert's features expectation is linear, the set of distributions over trajectories $P(\tau)$ depends exponentially on the inner product of reward parameters θ. Furthermore, high reward trajectories are exponentially more likely to be sampled from an expert than low reward trajectories [135]. In this context, to learn the expert's reward function, we maximize the log likelihood of the observed demonstrations under the distribution derived with the maximum entropy principle. Within the log likelihood, the product of the trajectories is treated as the sum of the conditional probabilities of all possible expert trajectories given the parameter θ, $\theta^* = \underset{\theta}{\operatorname{argmax}} L(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{\tau_i \in \mathcal{D}} \log P\left(\tau_i \mid \theta\right)$. Gradient descent is then used to maximize the likelihood of the expert's trajectories. Since the model of the environment is known, and the reward function is assumed to be linear in the features, the gradient consists of two main components (Equation 2.3.1). The first component is the empirical estimate of the expert's features, the second component is the sum over all states and depends on the state visitation frequency under the current policy. Dynamic programming can be used to efficiently calculate this *visitation frequency measure* (Equation 2.3.1).

$$\begin{aligned} \nabla_\theta L(\theta) = \tfrac{1}{m} \sum_{\tau_i \in \mathcal{D}} & \underbrace{\mu\left(\tau_i\right)}_{\text{expert state features}} - \sum_s \underbrace{d_s^\theta}_{\text{state occupancy measure}} \phi(s) \\ & = \tfrac{1}{T} \sum_{s' \in \tau_i} \phi\left(s'\right) \end{aligned} \tag{2.2}$$

$$d_{t+1,s'} = \sum_a \sum_s d_{t,s} \pi_\theta(a \mid s) P\left(s' \mid s, a\right) \tag{2.3}$$

As previously mentioned, having a complete set of demonstrations that cover the entire state space is often impractical. Research efforts in active learning for reward estimation in inverse reinforcement learning try to deal with this problem by allowing the agent to query the demonstrator in specific states, thereby gaining the ability to choose best situations to be demonstrated and requiring less demonstrations [136].

Interactive Direct Policy Learning

Interactive direct policy learning (DPL) considers situations in which learning is supervised through the use of reward signals in response to the observed outcomes of actions [137, 21]. In practise, Interactive DPL involves collecting the expert's demonstrations, applying supervised

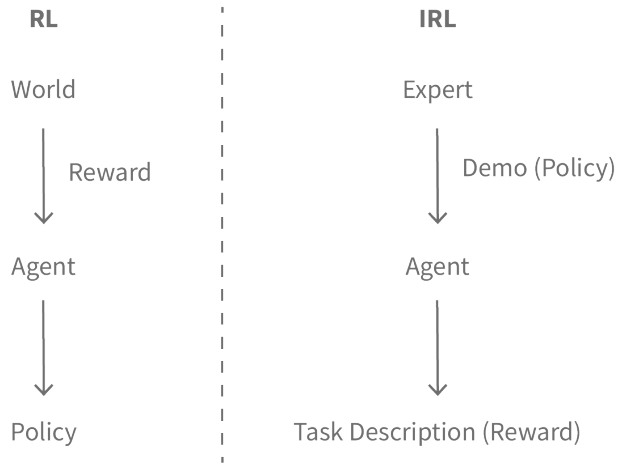

Figure 2.5: Comparison between Reinforcement Learning and Inverse Reinforcement Learning

learning to define a policy, rolling out that policy in the environment, and than receive feedback from the demonstrator about the roll out trajectories, thus receive additional training data to improve the learned policy. This process can be thought as a generalization of behavioral cloning and analyzed through the lens of learning reductions, i.e., reducing an harder learning problem to an easier one, as supervised learning [138] [139]. Interactive direct policy learning methods construct a series of distributions, or supervised learning problems, that ideally converge to training the best policy in our policy class for the general imitation learning problem. Starting with behavioral cloning, interactive direct policy learning enriches the learning process by recurring to an oracle to provide feedback on the state distributions by the ongoing policy rather than only by the expert's demonstrations policy. This means that it is possible to train the agent for any possible state and compare the agent's policy with the expert's policy in specific states. The information about what the expert would have done in any state can be aggregated, i.e., data aggregation (DAgger) [140], or policy aggregation (SEARN, SMILe) [141, 138], and used for training. By training a single policy on the union of the collected data sets and distributions, the agent learns from all the mistakes that has been done in every round of training. Alternatively, by training a single policy on the current distributions of states and mistakes, policy aggregation combines geometrically [7] all the previous policies.

Other approaches that aim to integrate human feedback in reinforcement learning include the TAMER framework proposed by [21], and the COACH framework proposed by [142]. In TAMER, humans provide interactive numeric feedback as the agent takes action. The agent estimates a target reward function by interpreting the human feedback as exemplars of this function. In COACH, human feedback is policy dependent and and treated as advantage signals. In both

[7] geometric aggregation or blending follows the geometric progression, therefore every term bears a constant ratio to its preceding term.

TAMER and COACH, the agent passively receives critiques without actively querying to address the ambiguity in reward learning. In contrast, there has been significant work on active IRL approaches. Most active IRL algorithms use a Bayesian approach [130].

To summarize, in behavioral cloning (BC) and interactive policy learning, the agent learns directly the policy from the demonstrations. Whereas in inverse reinforcement learning (IRL) we are learning the reward function from demonstrations and then indirectly learning the policy by maximizing the reward in a reinforcement learning subroutine. BC does not require access to the environment, as opposed to interactive policy learning and inverse reinforcement learning which require access to the environment to roll out a policy and then collect feedback. Only interactive policy learning requires an interactive demonstrator to be available during training, while the other two approaches can operate entirely on pre-collected demonstrations.

2.3 Explanation-Based Learning

The idea of learning from explanations share many commonalities with approaches to acquire knowledge from other expert agents. Over three decades ago scholars already began to investigate how explanation-based approaches could be applied to systems that are based on high-level symbolic (human-readable) representations of problems (i.e., symbolic-AI systems) [143]. Lewis [144] refers to analysis-based generalization methods to group work on explanation-based learning, and analogical generalization. In contrast to inductive approaches that examine a number of examples of a to-be-learned concept and construct a condensed description that is satisfied by all the examples, analysis-based generalization methods discern the essential features of a single example to explain what makes this example an example and find the class of examples for which the same explanation holds.

In Explanation-based learning (EBL) and generalization (EBG), a specific problem's solution is framed into a form that can be later used to solve conceptually similar problems [145]. EBL serves as a method to generalize iterative or recursive processes [146], chunk patterns [147], operationalize acquired rules [148], and reason about analogies [149]. EBL has found application in a wide range of areas, such as planning [150, 151], and natural language processing [152]. It is interesting to notice that even preliminary work on the topic refers to the notion of causality. In Mooney et al. [153] the basic task of an *understander*, i.e., learner, is to construct a causally complete representation called *model*.

The use of explanations to generalize better from fewer examples have been also explored in artificial neural nets. A promising approach is the integration of explanations as prior knowledge encoded in previously learned neural networks. Explanations provide to the learner with a

structure to interpret the observed example and infer additional information about the shape, or slope, of a target reward function [154]. Explanation-based neural network learning (EBNN) is a neural network analogue to symbolic explanation-based learning methods (EBL). It extends explanation-based learning to cover situations in which prior knowledge is approximate and is itself learned from scratch. Within EBNN, the need for large training data sets is replaced with previously learned domain theory, represented by neural networks.

2.3 Learning Rewards from Explanations

In human-in-the-loop machine learning, explanations generally take the form of verbal instructions. In the context of reinforcement learning and inverse reinforcement learning particular emphasis has been placed on incorporating verbal instructions by either combining natural language with demonstrations [155] and by using sentiment analysis to filter natural language input into advice [60, 156].

The work of Babes-Vroman et al. [155] introduces an architecture for sentence–trajectory pairs, where the learner has access to natural language input, and demonstrations of appropriate behavior. Their architecture include maximum likelihood inverse reinforcement learning to estimate the expert's reward function from linguistic feedback available at different stages in the learning process.

Accounting for the fact that humans often do not specify state information in their advice, e.g., *Mario should jump on enemies*, the work of Krening et al. [60] propose a method named *object-focused advice* in which the human advice is tied to objects instead of specific states and is generalized over the objects' state space. Their experiment collects information on the nature of explanations, the accuracy of their sentiment analysis to filter explanations, and the performance of agents trained with *object-focused advice*. Their results show that their method is able to capture human explanations without state information and increases the performance of reinforcement learning agents. They observe that free-form explanations, i.e., human explanations not constrained to a template, vary in many ways. The level of detail and abstraction used to describe desired actions seem to reflect the amount of prior knowledge the learning agent is assumed to have. However, while a sentiment filter can process free-form explanations, the majority of the sentences are not actionable and cannot be directly utilized as advice.

In line with the use of sentiment analysis to filter linguistic feedback, in the work of Sumers et al. [156] the learner grounds linguistic feedback to elements of the task, e.g., *"Good job"* refers to prior behavior, whereas *"You should have gone to the living room"* refers to an action,

and assigns a positive or negative sentiment to that behavior. The positive or negative valence of the behavior implies a positive or negative rewards on its features. Linguistic feedback are processed as *evaluative*, *imperative*, and *descriptive* feedback. An evaluative feedback corresponds to a scalar value in response to the agent's actions and have positive or negative valence (+1/-1). An imperative feedback gives information about the correct action in a given state by mapping the language input into a set of state-action pairs. A descriptive feedback provides information about the state transition function, i.e., how the teacher's preference changes in response to an action. Their results show that the learner is able to learn from all types of linguistic feedback, obtaining the best scores when trained with descriptive feedback.

The use of linguistic feedback as teaching signals to transfer knowledge has been studied both in the context of human social learning and machine learning. Explanations help establish a connection between what has been observed and its causes, and serve as a principled basis for generalization [29]. Consequently, explanations scaffold causal learning and have a crucial role in inference [157]. Following this idea, our work also generate explanations in the form of sentence-trajectories and uses maximum likelihood inverse reinforcement learning to find a weighting of the state features that (locally) maximizes the probability of these trajectories. We provide a framework to learn from explanations and allow a fair comparison with other types of teaching signals, i.e., reward, demonstration. In addition, we evaluate the generated teaching signals in a user study, accounting for different situations and positions of the learner with respect to the goal.

2.4 Effects of Agent's Explanations on Teamwork

Group dynamics often play a role in human social learning [158, 159]. The act of collaboration (i.e., *the act of working with another person or group of people to create or produce something*) and cooperation (i.e., *the fact of doing something together or of working together towards a shared aim*) in group interactions is not only interesting for researchers in the area of human-machine interaction but is also widely studied by social sciences to obtain knowledge on how cooperation can be manipulated. In particular, to understand how to boost individuals contribute to a public good [160]. Several studies, both theoretically and empirically, shown that explainability has a positive effect on cooperation. For instance, Fudenberg and Maskin [161] demonstrated that explainability of past choices by the group members is necessary to maintain a sustainable and stable cooperation. Davis et al. [162] shown that explainability allows cooperative players to indicate their cooperative intentions, which may induce others to similar cooperative behaviors.

Explainability is often associated with appropriate mutual understanding and trust that leads to effective collaboration between agents [76, 110, 97, 58, 115, 163]. Trust appears as a common measure to assess the effect of explainability and it is related to the level of observability, predictability, adjustability, and controllability, as well as mutual recognition of common objectives between the system and its user.

Enabling and facilitating bi-directional intent recognition has been identified as a main challenge to successful collaboration between humans and AI systems. Research in this direction focuses on pre-execution communication and legible motion.

Studies on legibility in multi-party interactions suggest that improving the group's average legibility improves the group's general understanding of a robot's intention, thus improving team efficiency and safety in the interaction [164].

Although there are several studies on how explainability affect human-AI teamwork, existing work focuses mainly on complementary team performance (CTP), i.e., a level of performance beyond the ones that can be reached by AI or humans individually, while neglecting the effect of explainability on other outcomes of the human-AI collaboration.

Contrary to what we could assume, collaborative scenarios can also encourage anti-collaborative practices that derive from the fact that group members rely on the contribution of others and therefore invest less in their actions , i.e., *free riding*. For this reason, combining the investigation of the behavioral model of the contributors in relation to the different strategies of the team members and the transparency of the decision-making process of the contributors turns out to be useful for the design of systems that aim to facilitate and foster collaboration.

To the best of our knowledge, there is no evidence regarding the role of explainability in combination with the agent's strategies on group dynamics, e.g., group identification, trust etc.. We investigate this aspect by implementing an explainability module in a public goods game and measuring the effect of the revealing the strategy of intelligent agents on human collaborative choices.

Chapter 3

Explainable Embodied Agents Through Social Cues

The issue of how to make embodied agents explainable has experienced a surge of interest over the last three years, and, there are many terms that refer to this concept, e.g., transparency or legibility. One reason for this high variance in terminology is the unique array of social cues that embodied agents can access in contrast to that accessed by non-embodied agents. Another reason is that different authors use these terms in different ways. This chapter reviews the existing literature on explainability and organize it by (1) providing an overview of existing definitions, (2) showing how explainability is implemented and how it exploits different social cues, and (3) showing how the impact of explainability is measured. Additionally, the chapter lists open questions and challenges that highlight areas that require further investigation by the community. This provides the interested reader with an overview of the current state-of-the-art.

Research Questions

We confined the scope of this review to embodied agents and tried to answer the following questions:

- What is the definition of explainability?
- How is explanability implemented?
- How social cues are exploited in explainable embodied agents?
- How is explainability measured?
- Which open questions and challenges have been identified by the community?

Methods

For this review, we chose to use a keyword based search in Scopus[1] database to identify relevant literature, as this method makes our search reproducible.

First, we identified a set of relevant papers in an unstructured manner based on previous knowledge of the area. From each paper, we extracted both, the indexed and the author keywords, and rank ordered each keyword by occurrence. Using this method, we identified key search terms such as *human-robot interaction, transparent, interpretable, explainable,* or *planning*.

We then grouped these keywords by topic (more details about the query in [12]) and performed a pilot search on each topic to determine how many of the initially identified papers were recovered. We then combined each group using *AND*, which led to a corpus of 263 papers. All authors participated in this initial extraction process.

[1] https://www.scopus.com/

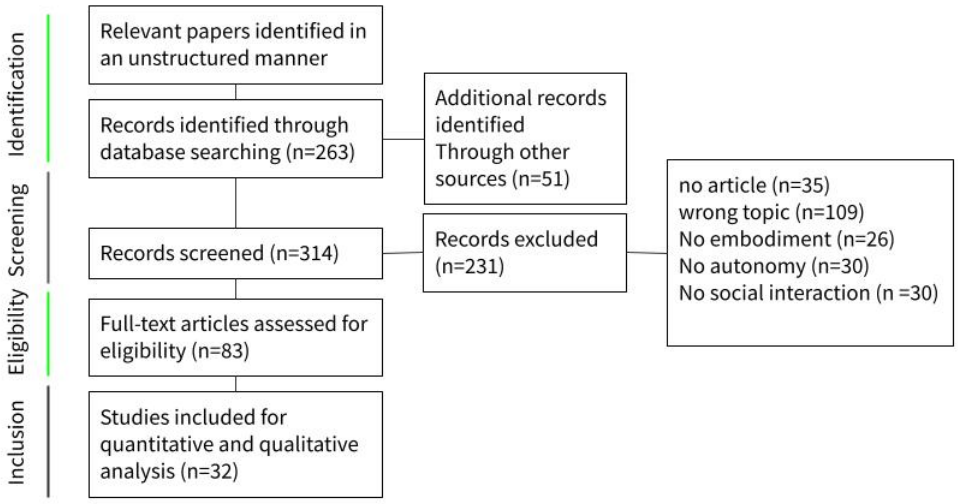

Figure 3.1: Flow diagram of study inclusion for the literature review.

Next, we manually filtered this list to further remove unrelated work by judging relevance based on titles, abstracts, and full text reads. To ensure selection reliability, both main authors rated inclusion of each paper independently. If both labelled the paper as relevant, we included the paper; similarly, if both labelled it as unrelated, we excluded it. For papers with differing decisions, we discussed their relevance and made a joint decision regarding the paper's inclusion. This left us with 32 papers for the final review.

For the excluded papers, each main author indicated why a paper was excluded for the following reasons:

- **no article** The paper was a book chapter or review paper. ($\sim 15.63\%$ excluded)
- **wrong topic** The paper presented work in a different focus area, e.g, material science, teleoperation, or generically making robots more expressive (without considering explainability). ($\sim 45.98\%$ excluded)
- **wrong language** The paper was not written in English. ($\sim 0.46\%$ excluded)
- **no embodiment** The paper did not consider an embodied agent. ($\sim 11.26\%$ excluded)
- **no autonomy** The paper did not consider autonomous embodied agents. ($\sim 13.33\%$ excluded)
- **no social interaction** The paper did not investigate explainability in a context where a human was present. ($\sim 13.10\%$ excluded)

Table 3.1: Papers on Explainability Ordered by Social Cues

Category	Paper
Speech	[165, 108, 166, 114, 100]
Text	[88, 89, 91, 94, 167, 99, 58, 116, 98, 168, 113, 169]
Movement	[106, 93, 76, 102, 109, 108, 166, 115, 98, 170]
Imagery	[171, 106, 110, 166, 114, 104]
Other/Unspecified	[101, 166]/ [97, 172, 173, 105]*

3.1 Definition of Explainability in Embodied Agents

Our Definition

After reviewing existing definitions we provide a definition that aims to be comprehensive for our literature. **We define the explainability of embodied social agents as their ability to provide information about their inner workings using social cues, such that an observer (target user) can infer how/why the embodied agent behaves the way it does.**

Social Cues

We have claimed above that embodied agents can become explainable using unique types of social cues that are not available to agents lacking such embodiment. One example is the ability to point to important objects in a scene - assuming the agent has an extremity that affords pointing. A non-embodied agent has to use a different way to communicate the importance of that object.

Hence, we screened the core papers to check which modality the authors deployed to make the agent more explainable. We then logically grouped the core papers based on these types of social cues and identified five groups:

- **Speech** A lexical statement uttered verbally using a text-to-speech mechanism.
- **Text** A lexical statement displayed as a string presented as an element of a typically screen-based user interface.
- **Movement** A movement that is either purely communicative, or that alters an existing movement to make it more communicative.
- **Imagery** A drawing or image (often annotated) presented as an element of a user interface (typically screen-based).

- **Other/Unspecified** All papers that use social cues that do not fit within above set of categories, or where the authors did not explicitly specify the modality (the latter is marked with an asterisk*).

This grouping is shown in table 3.1. Surprisingly, our search did not yield any papers that investigate non-lexical utterances (beeping noise, prosody, etc.), which was contrary to our expectations. A possible explanation for this could be that our search terms did not capture a broad enough scope, because experiments investigating such utterances may use yet again a different terminology. Another possibility could be that it seems much harder to communicate an explanation through *beeps and boops* instead of using speech; especially when considering the wide availability of text-to-speech synthesizers (TSSs).

The wide availability of TTS synthesizers may also explain another interesting result of this analysis. Many works focus on lexical utterances (speech and text). Potentially, such utterances are seen as easier to work with when giving explanations because of the high expressivity of natural language.

On the other hand, lexical utterances may add additional complexity to the interaction, because a sentence has to be interpreted and understood, whereas, other social cues may be faster/easier to interpret. While there does exist work that investigates the added cognitive load of having explainability versus not having explainability [113], comparing lexical utterances with other forms of explainability is currently underexplored. This prompts the question of whether lexical utterances are always superior to achieve explainability and, if not, under what circumstances other social cues perform better.

Evaluation Methods

Existing works assess the effects of explainability on a variety of measures including but not limited to, self-reported understanding of the agent [99], number of successful task completions [113], number of false decisions [113], task completion time [93], number of irredeemable mistakes [100] or trust in automation [171]. During our review three major categories of measurements emerged:

- **Trust** measures how willing a user is to agree with a decision of a robot - based on the provided explanation -, how confident a user is about the embodied agent's internal workings (internal state), or if the user agrees with the plan provided by the robot (intent). It is measured using a self-report scale.

- **Robustness** measures the avoidance of failure during the interaction. Typically researchers

want to determine whether the embodied agent's intent has been communicated correctly. It is often measured observationally, e.g., by counting the frequency of successful achievements of a goal.

- **Efficiency** measures how quickly the task is completed. The common hypothesis behind using this measure is that the user can adapt better to a more explainable robot, and form a more efficient team. It is commonly measured by wall clock time, or number of steps until the goal.

Among these measures, trust received the most attention. While there is large variance in which scale is used (often scales are self-made), a common element in the studies is the use of self-report questionnaires.

Although the consensus is that the presence of explainability generally increases trust (see Table 3.2), how effective a particular social cue is in doing so has received much less attention. Comparisons that do exist often fail to find a significant difference between them [113, 171]. Similarly, due to the large range of mechanisms tested - and the even larger array of scenarios -, there is little work on how robust a specific mechanism performs across multiple scenarios. Hence, while some form of explainability seems to be clearly better then none, which specific mechanism to choose for which specific situation remains an open question.

Less studied, but no less important, is the effect of explainability on the robustness of an interaction. Research on the interplay between explainability and robustness uses tasks where mistakes are possible, and measures how often these mistakes occur [171, 166]. The core idea is that participants create better mental models of the robot when it is using explainability mechanisms. Better models will lead to better predictions of the embodied agent's future behaviour, allowing participants to anticipate low performance of the robot, and to avoid mistakes in task execution. However, experimental evidence on this hypothesis is not always congruent, with the majority of studies showing support for the idea, e.g., [166], and other studies finding no significant difference, e.g., [171]. As the majority does find a positive effect, we can conclude that explainability does help improve reliability, although not in all circumstances. A more detailed account of when it does or does not remains a subject for future experimental work.

Finally, efficiency is a metric that some researchers have considered while manipulating explainability. It has been operationalized by comparing wall clock time until task completion across conditions [93], or time until human response [91]. Of the three types of measures, this type has received the least attention, and the findings are quite mixed. Approximately half of the analysed papers find that making embodied agents explainable makes the team more efficient, while the other half find no difference. However, a clear explanation for these conflicting findings

Table 3.2: Papers on Explainability by Measure

Type	Outcome	Papers
Robustness	Positive	[91, 169, 113, 102, 115, 109, 76, 166, 174, 105]
Robustness	Negative	
Robustness	Non-significant	[93]
Robustness	No statistical test	[110, 98, 175]
Trust	Positive	[93, 91, 176, 171, 174, 170, 168, 169, 104]
Trust	Negative	
Trust	Non-significant	[89]
Trust	No statistical test	[88, 97, 102, 100, 116, 94, 167, 99]
Efficiency	Positive	[176, 113, 89]
Efficiency	Negative	
Efficiency	Non-significant	[166, 168, 91, 93]
Efficiency	No statistical test	[114, 97]
other	any	[108, 107, 177, 172, 58]

remains a topic of future work.

Table 3.2 shows the core papers grouped by the evaluation methods discussed above and indicates whether the effect of explainability on it was positive, negative, or non-significant. One important note is that many papers introduce a measurement called *accuracy*; however, usage of this term differs between authors. For example, Chao et al. [93] used accuracy to refer to the embodied agent's performance after a teaching interaction; hence it was being a measure of robustness, whereas Baraka and Veloso [101]'s accuracy referred to people's self-rated ability to predict the robot's move correctly, a measure of trust.

In summary, there is enough evidence that explainability offers a clear benefit to virtual embodied agents in building trust, with some support for physical embodied agents. Additionally, there is evidence that explainability can decrease the chance of an unsuccessful interaction (improve robustness). However, papers looking to improve the efficiency of the interaction find mixed results. A possible explanation for this could be that while explainability makes the interaction more robust, the time added for the embodied agents to display and for the human to digest the additional information nullifies the gain in efficiency.

In addition to the above analysis, this section identified the following open questions: (1) Is a particular explainability mechanism best suited for a specific type of embodied agent, a specific type of scenario, or both? (2) What are good objective measures with which we can measure trust in the context of explainability? (3) Why does explainability have a mixed impact on the efficiency of the interaction?

3.2 Findings

In the above sections we provided a focused view on four key aspects of the field: (1) definitions used, and the large diversity thereof, (2) which social cues and (3) algorithms are used to link explainability mechanisms to the embodied agent's state or intent, and (4) the measurements to assess explainability mechanisms. What is missing is a discussion of how these aspects relate to each other when looked at from a $10,000$ foot view, and a discussion of the limitations of our work.

It is almost self-explanatory that the scenario chosen to study a certain explainability mechanism depends on the author's research goal. As such, it is unsurprising that we can find a large diversity of tasks, starting from evaluation in pure simulation [58], or discussions of hypothetical scenarios [176, 116] all the way to joint furniture assembly [114].

3.2 Human Decision-Making

The most dominant strand of research has its origin in decision making, and mainly views the robot as a support for human decisions [88, 89, 91, 94, 167, 171, 169, 113, 168]. In this line of research, explainability is mostly commonly defined via the SAT-model (i.e., Situation Awareness-Based Agent) [178]. One of the key questions is how much a person will trust the embodied agent's suggestions, based on how detailed the given justification for the embodied agent's decision is. While these studies generally test a virtual agent shaped like a robot, the findings here can be easily generalized to the field of human-computer-interaction (HCI), due to their design. Hence, SAT model-based explanations can help foster trust not only in HRI, but also in the domain of expert systems and AI. Hence, this work partially overlaps with the domain of explainable AI (XAI).

3.2 System's Robustness

The second strand of research sets itself apart by using humans as pure observers [97, 99, 76, 98, 115, 174, 170, 175, 102, 117]. Common scenarios focus on communicating the embodied agent's internal state or intent by having humans observe a physical robot [175, 115, 174] or video recordings/simulations of them [170, 174, 102]. Other researchers choose to show maps of plans generated by the robot and explanations thereof [97, 99, 76, 98]; the researchers' aim here is to communicate the robot's intent. In all scenarios, the goal is typically to improve robustness, although other measures have been tested.

Particularly well done here is the work of Baraka et al. [174], who first describe how to enhance a robot with LED lights to display its internal state, use crowd sourcing to generate expressive patterns for the LEDs, and then validate the pattern's utility in both a virtual and a physical user study. This pattern of having participants - typically from Amazon Mechanical Turk (AMT) - generate expressive patterns in a first survey, and then validate them in a follow-up study was also employed by Sheikholeslami et al. [175] in a pick-and-place scenario. We think that this crowdsourcing approach deserves special attention, as it will likely lead to a larger diversity of candidate patterns compared to an individual researcher generating them. Considering the wide availability of online platforms, such as AMT and Prolific, this is a tool that future researchers should leverage.

3.2 Human-Robot Interaction

A third strand of research investigates explainability in interaction between a human and a robot [104, 109, 114, 166, 100, 108, 93] or a human and an AI system [110]. Studies in this strand investigate the impact of different explainability mechanisms on various interaction scenarios and whether they are still useful when the human-robot dyad is given a concrete task. This is important, because users can focus their full attention on the explainable behaviour in the observer setting; in interaction scenarios, on the other hand, they have to divide their attention. Research in this strand is more heterogeneous, likely due to the increased design complexity of an interaction scenario. At the same time, the amount of research done, i.e., the number of papers identified, is less than the research done following the observational design above; probably because of the the above mentioned added complexity. Nevertheless, we argue that more work on this strand is needed, as we consider testing explainability mechanisms in an interaction as the gold standard for determining their utility and effectiveness.

Finally, some researchers examined participants' responses to hypothetical scenarios [176, 116]. The procedure in these studies is to first describe a scenario to participants in which a robot uses an explainability mechanism during an interaction with a human. Then, participants are asked to give their opinion about this interaction, which is used to determine the utility of the mechanism. This method can be very useful during the early design stages of an interaction, and can help find potential flaws in the design before spending much time implementing them on a robot. At the same time, it may be a less optimal choice for the final evaluation, especially when compared to the other methods presented above.

3.2 Challenges in Explainability Research

Shifting the focus to how results are reported in research papers on explainability, we would like to address two challenges we faced while aggregating the data for this review.

The first challenge is the large diversity and inconsistency of language used in the field. Transparency, explainability, expressivity, understandability, predictability and communicability are just a few examples of words used to describe explainability mechanisms. Authors frequently introduce their own terminology when addressing the problem of explainability. While this might allow for a very nuanced differentiation between works, it becomes challenging to properly index all the work done, not only because different authors addressing the same idea may use different terminology but also especially because different authors addressing different ideas end up using the same terminology.

Other reviews on the topic have pointed this out as well [117, 179], and it became a challenge in our review, as we cannot ensure completeness of a keyword search based approach. The most likely cause of this is because the field is seeing rapid growth, and precise terminology is still developing.

This work tries to address this first challenge by showing how different terms are used to identify similar concepts and providing a definition that aims to be comprehensive for the surveyed papers.

The second challenge was that many authors only define the explainability mechanism they investigate implicitly. We often had to refer to the concrete experimental design to infer which mechanism was studied. While all the important information is still present in each paper, we think that explicitly stating the explainability mechanism under study can help discourse regarding explainability become much more concrete.

In extension, some authors have implemented explainability mechanisms on robotic systems that are capable of adapting their behaviour or performing some kind of learning. In many cases, these learning algorithms were unique implementations, or variations of standard algorithms, e.g., reinforcement learning, which make them very interesting. How to best incorporate an explainability mechanism into such a framework is still an open question. Unfortunately, we found that the details of the method are often underreported and that we could not extract enough data on what has been done so far. We understand that this aspect is often not the core contribution of a paper and that space is a constraint. Nevertheless, we would like to encourage future contributions to put more emphasis on how explainability mechanisms are integrated into existing learning frameworks. Technical contributions such as this could prove very valuable for defining a standardized approach to achieve explainability using embodied social agents.

Open Questions

While performing the review, we identified a set of open questions. For convenience we enumerate them here:

1. What are good models to predict/track human expectations/beliefs about the embodied agent's goals and actions?
2. What are efficient learning mechanisms to include the human in the loop when building explainability into embodied agents?
3. How does the environment and embodiment influence the choice of social cues used for explainability?
4. What are good objective measures by which we can measure trust in the context of explainability?
5. Why does explainability not have a strictly positive impact on the efficiency of the interaction?

Table 3.3: Identified Categories by Paper

CitationKey	Definition					Social Cues					Measurement				Learning Paradigm
	transparency	explainability	expressivity	other	none	Movement	Text	Speech	Imagery	other/None	Trust	Robustness	Efficiency	Other	ML
Akash et al. [88]	X	.	.	.	.	.	X	.	.	.	X	.	.	.	.
Akash et al. [89]	X	.	.	.	.	.	X	.	.	.	X	.	X	.	.
Baraka and Veloso [101]	X	.	X	.	.	.	.	.	.	X	X	X	.	.	.
Boyce et al. [171]	X	.	.	.	.	.	.	.	X	.	X	.	.	.	.
Brown and Laurier [106]	.	.	.	.	X	X	.	.	X	.	.	.	.	X	.
Chakraborti et al. [97]	.	X	.	.	.	.	.	.	.	X	X	.	X	.	.
Chao et al. [93]	X	.	.	.	.	X	.	.	.	.	X	X	X	.	X
Chen et al. [91]	X	.	.	.	.	.	X	.	.	.	X	X	X	.	.
Fischer et al. [165]	X	.	.	.	.	.	.	X	.	.	X	.	.	.	.
Floyd and Aha [94]	X	.	.	.	.	.	X	.	.	.	X	.	.	.	.
Floyd and Aha [167]	X	X	.	.	.	.	X	.	.	.	X	.	.	.	.
Gong and Zhang [99]	.	X	.	.	.	.	X	.	.	.	X	.	.	.	X
Hayes and Shah [58]	X	X	.	.	.	.	X	.	.	.	.	.	.	X	X
Huang et al. [76]	.	.	.	X	.	X	.	.	.	.	.	.	.	.	X
Khoramshahi and Billard [172]	.	.	.	.	.	.	.	.	.	X	.	.	.	.	X
Kwon et al. [102]	.	.	X	.	.	X	.	.	.	.	X	X	.	.	.
Lamb et al. [109]	.	.	.	.	X	X	.	.	.	.	.	X	.	.	.
Lee [173]	.	.	.	X	.	.	.	.	.	X	X	.	.	.	.

Table 3.3 continued from previous page

CitationKey	Definition					Social Cues					Measurement				Learning Paradigm
	transparency	explainability	expressivity	other	none	Movement	Text	Speech	Imagery	other/None	Trust	Robustness	Efficiency	Other	ML
Legg et al. [110]	.	.	.	.	X	.	.	.	X	.	.	X	.	.	X
Lutkebohle et al. [108]	.	.	.	.	X	X	.	X	.	.	.	.	.	X	X
Perlmutter et al. [166]	X	.	.	.	.	X	.	X	X	X	.	X	X	.	X
Poulsen et al. [116]	X	.	.	.	.	.	X	.	.	.	X	.	.	.	.
Roncone et al. [114]	X	.	.	.	.	.	.	X	X	.	.	.	X	.	.
Schaefer et al. [104]	.	.	.	X	.	.	.	.	X	.	X	.	.	.	.
Sciutti et al. [115]	.	.	.	X	.	X	.	.	.	.	.	X	.	.	.
Sheikholeslami et al. [175]	.	.	.	X	.	X	.	.	.	.	.	X	X	.	.
Sreedharan et al. [98]	.	X	.	.	.	.	X	.	.	.	.	X	.	.	.
Tabrez et al. [85]	.	X	.	.	.	.	.	X	.	.	X	.	.	.	X
Wang et al. [168]	.	X	.	.	.	.	X	.	.	.	X	.	X	.	.
Wang et al. [113]	.	X	.	.	.	.	X	.	.	.	.	X	X	.	.
Wang et al. [169]	.	X	.	.	.	.	X	.	.	.	X	X	.	.	X
Zhou et al. [170]	.	.	X	.	.	X	.	.	.	.	X	.	.	.	.
Grigore et al. [105]	.	.	.	X	.	.	.	.	.	X	.	X	.	.	.

Chapter 4

Effects of Agents' Explanations on Teamwork

Despite the interest in explainability, there are still few experimental studies on how explainability affects teamwork, in particular in collaborative situations where the strategies of others, including agents, may seem obscure. We explored this problem using a collaborative game scenario with a mixed human-agent team. We investigated the role of explainability in the agents' decisions, by having agents that reveal and tell the strategies they adopt in the game, in a manner that makes their decisions transparent to the other team members. The game embraces a social dilemma where a human player can choose to contribute to the goal of the team (cooperate) or act selfishly in the interest of his or her individual goal (defect). We designed a between-subjects experimental study, with different conditions, manipulating the explainability in a team. The results show an interaction effect between the agents' strategy and explainability on trust, group identification and human-likeness. Our results suggest that explainability has a positive effect in terms of people's perception of trust, group identification and human likeness when the agents use a Tit for Tat or a more Individualistic strategy. In fact, adding transparent behaviour to an unconditional cooperator negatively affects the measured dimensions.

Research Questions

With this work we investigated the following research questions:

- Do agent's explanations have an affect on the perception of intelligent agents during human-agent teamwork?
- Do agent's explanations affect humans' pro-social behaviors?

4.1 For the Record Game Platform

Game Scenario

For this research we used the game "For the Record". "For the Record" is a public goods game that embraces a social dilemma where a human player can choose to contribute to the goal of the team (cooperate) or act selfishly in the interest of his or her individual goal (defect). In linear public goods environments *maximizers have a dominant strategy to either contribute all of their tokens or none of their tokens to a group activity* [180, 181]. In the "For The Record" experimental scenario, three players, one human, and two artificial agents, have the goal of publishing as many albums as possible. The number of albums to be created and produced matches the number of rounds to play, in our case, 5 rounds and if players fail 3 albums they lose the game. During the first round, each player starts playing by choosing the preferred instrument that can be used to

create the album. Starting from the second round each player has two possible actions and they concern the possibility of investing in the instrument's ability (contributing to the success of the album) or in the marketing's ability (contributing to the individual monetary value, or personal profit, obtained after the album's success). This investment is translated into the number of dice that the player can use, in the first case to play the instrument and helps to create the album, while in the second case to receive profit. During the creation of the album, each player will contribute equally to the value obtained from the roll of the dice, and the number of die available to the player will depend on the level/value of the skill (marketing or instrument). The score of an album consists of adding up the values achieved by each player during his performance. After creating the album, the band has to release it on the market. The market value is evaluated by rolling 2 dice of 20 faces. If the market value is higher than the album score, than the album is considered a "Fail". On the other hand, if the market value is less than or equal to the score on the album, that album is considered a "Mega-hit". From the fourth round on, the band enters the international market, which means that the market value is evaluated by rolling 3 die of 20 faces (instead of the 2 previous dices). This increases the difficulty of getting successful albums. The game has always been manipulated to return a victory.

4.2 Experimental Design

The objective of this study was to investigate the effect of the explainability and strategy of virtual agents on human pro-social behavior in a collaborative game. Despite having hypothesized that explainability would affect several measures of teamwork, we have also manipulated the agents' strategy to confirm if the results would provide similarly when the agents adopted different strategies. In a two by three (2 x 3) between-subjects design, resulting in six experimental conditions, we manipulated the agents' explainability and the agents' strategy, respectively. The two levels of explainability were:

- **Transparent**: The agents explain their strategy;
- **Non-transparent**: The agents do not explain their strategy.

The three possible strategies for the agents were:

- **Cooperative**: The agents always cooperate;
- **Individualistic**: The agents cooperate only if the last round has been lost;
- **Tit for Tat**: The agents cooperate only if the player cooperate.

Hypotheses

We expected that the explainability of the agents will positively affect teamwork and make the agents' strategy easily to interpret. We also expected explainability to increase trust and facilitate collaboration due to mutual understanding and shared responsibilities. Therefore we have the following hypotheses:

- H1: The agents' explainability increases the number of Cooperative choices of the human player;
- H2: The agents' explainability results in greater trust and group identification;
- H3: The agents' explainability increases the likeability and human likeness of the artificial player;

The hypothesis that the explainability increases the number of cooperative choices is based on the fact that transparency about choices tends to lead to an increase in contributions and collusion [160]. The hypothesis that positive effect of explainability on trust and group identification relies on the evidence that explainability have the (perhaps counter-intuitive) quality of improving operators' trust in less reliable autonomy. Revealing situations where the agent has high levels of uncertainty develops trust in the ability of the agent to know its limitations [182, 183, 184, 185]. The hypothesis that the agents' explainability results in greater likeability and perceived human likeness of the artificial player refers to the experimental evidence of Herlocker et al. showing that explanations can improve the acceptance of automated collaborative filtering (ACF) systems [186].

We conducted a between-subject user study using the Mechanical Turk and the "For The Record" game [187].

Agents' Explainability Manipulation

The interactive agents commented some game events through text in speech bubbles, e.g., *That was very lucky!* or *Lets record a new album*.

The duration of such stimuli depend on the number of words shown, according to the average reading speed of 200-250 words per minute. However, the speech bubbles containing the manipulation of each experimental condition lasted twice as much to make sure the participants would read them (Fig. 4.1).

Table 1 shows the explanation given by the artificial agents while they are choosing the main action of adding a point to either the instrument or the marketing in the transparent and

Figure 4.1: Example of a speech bubble with the explanation of the agents' strategy.

non-transparent conditions:

Strategy \ Explainability	**Explainable**	**Non-explainable**
Cooperative	*1. "My strategy is to always improve the instrument."* *2. "My plan is to always improve the instrument."*	*1. "I am going improve the [instrument/ marketing]."* *2. "I will put one more point on my [instrument/marketing]."*
Individualistic	*1. "My plan is to improve my marketing skill only when the album success."* *2. "My plan is to improve my instrument skill only when the album fails."*	
Tit for Tat	*1. "My strategy is to improve my instrument skill only when you also improve your instrument."* *2. "My strategy is to improve my marketing skill only when you also improve your marketing."*	

Table 4.1: Manipulation of explainable and non-explainable behaviour for each agents' strategy

In the non-transparent conditions the agents explain what they are doing for that current round, in the transparent conditions they explicitly refer to their plans and intentions.

Metrics and Data Collection

To test our hypotheses and, therefore, analyse the effects of the strategy and explainability adopted by the agents, we used different metrics and items from standardized questionnaires. The self-assessed questionnaire included some demographic questions (e.g., age, gender and ethnicity), a single-item on their self-perceived competitiveness level, two items regarding the naturalness and human-likeness of the agents' strategies, and two validation questions to evaluate

the understanding on the rules of the game. The remaining measures are detailed as follows.

Cooperation Rate

The cooperation rate was an objective measure assessed during the game-play. In the beginning of each round, each player has to choose between to cooperate with the team (i.e., by upgrading the instrument skill) or to defect for individual profit (i.e., by upgrading the marketing skill). This measure sums up the total number of times the human player opted to cooperate and can range, in discrete numbers, from zero to four. It represents the degree of pro-sociality that the human participant expressed while teaming with the agents.

Group Trust

We chose the Trust items by Allen et al. in [188], which were explicitly designed for virtual collaboration to assess the trust through the agents. Trust is described as a key element of collaboration and is divided into seven items with a 7 points likert-scale from totally disagree to totally agree.

Multi-component Group Identification

Leach et al. identified a set of items for the assessment of the Group-Level Self-Definition and Self-Investment in [189]. The idea behind this scale is that individuals' membership in groups has relevant impact on humans behavior. Specifically designed items represent the five dimensions evaluated: individual self-stereotyping, in-group homogeneity, solidarity, satisfaction, and centrality. These items were presented with a Likert-type response scale that ranged from 1 (strongly disagree) to 7 (strongly agree). We decided to use the dimensions of homogeneity, solidarity and satisfaction as relevant metrics for our study.

Godspeed

The Godspeed scale was designed for evaluating the perception of key attributes in Human-Robot Interaction [190]. More precisely, the scale is meant to measure the level of anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. Each dimension has five or six items with semantic differentials couples that respondents are asked to evaluate in a 5 points Likert scale. We used the dimensions of the likeability (Dislike/Like, Unfriendly/Friendly, Unkind/Kind, Unpleasant/Pleasant, Awful/Nice) and perceived intelli-

gence (Incompetent/Competent, Ignorant/Knowledgeable, Irresponsible/Responsable, Unintelligent/Intelligent, Foolish/Sensible).

Participants

The participants involved in the study were 120, 20 participants per each experimental condition (Cooperative, Individualistic and Tit for Tat). Considering the study was done in MTurk and the fact that the experiment took more time than the turkers are used to, we introduced some attention and verification questions in order to ensure the quality of the data. The criteria to exclude participants were: not having completed the entire experiment; having reported an incorrect score of the game; and having provided wrong answers to the questions related to the game rules (e.g., *How many die are rolled for the international market?*). Consequently, we ran the data analysis on a sample of 80, 28 in the non-explainable conditions and 52 in the explainability conditions. The average age of the sample was 37 years (min = 22, max = 63, stdev = 8.78) and was composed of 52 males and 27 females and one other. The participants were randomly assigned to one of three condition of the strategy: 19 in the Cooperative condition (13 in the explainable condition and 6 in the non-explainable condition), 30 in the Individualistic condition (17 for the explainability condition and 13 in the non-explainable condition), 18 for the Tit for Tat condition (9 for the explainable condition and 11 in the non-explainable condition).

Procedure

Participants were asked to complete the task in around 40 minutes. The experiment was divided in three phases. The first phase consisted of the game tutorial, and lasted around 15 minutes. The second phase consists in playing a session of "For the Record" with the two artificial agents, which lasted around 15 minutes. The last phase was represented by the questionnaire and took round 10 minutes. We informed the participants about the confidentiality of the data, voluntary participation and the authorization for sharing the results with the purpose of analysis, research and dissemination. We specified that we were interested in how people perceive teamwork and the game strategies of the two artificial players they were going to play with. After finishing the experiment and providing their judgments, we thanked the participants for their participation giving them 4$.
We collected the data for the non transparent and the transparent condition separately, ensuring that none of the participants repeat the experiment twice.

Results

We analyzed the effects of our independent variables – Explainability (binary categorical variable *Explainable* and *Non-Explainable*) and strategy (three categories: *Cooperative, Individualistic* and *Tit for Tat*) – on the dependent variables.
The reliability analysis for the dimensions of the Trust scale, the Group Identification scale, the Godspeed scale as well as the Human likeness and Naturalness revealed excellent internal consistency among items of the same dimensions (Trust: $\alpha = 0.912$; Group Identification: $\alpha = 0.972$; Group Solidarity: $\alpha = 0.953$; Group Satisfaction: $\alpha = 0.969$; Group Homogeneity: $\alpha = 0.923$; Perceived Intelligence: $\alpha = 0.962$; Likeability: $\alpha = 0.978$; Human-likeness and Naturalness: $\alpha = 0.938$).

Cooperative Rate.

The analysis of the number of defects, revealed that the main effect of explainability was not significant ($F(1, 73) = 0.320, p = 0.573$), and the main effect of strategy was not significant ($F(3, 73) = 2.425, p = 0.072$). The interaction effect between the two factors was not significant ($F(2, 73) = 0.003, p = 0.997$). The specific values per each strategy were: Cooperative (M=1.11, SE=0.201, SD=875), Individualistic (M=1.70, SE= 0.153, SD=0.837), Tit for Tat (M=1.06, SE=0.249, SD=1.056).

Group Trust

The Analysis of Variance in Trust, showed that the main effect of explainability was not significant (F(1,73)=0.337, p = 0.563), and the main effect of strategy was significant ($F(3, 73) = 8.117, p < 0.001$). The specific values for each strategy were: Cooperative (M=5.25, SE=0.265, SD =1.154), Individualistic (M=4.42, SE=0.230, SD=1.261), Tit for Tat (M=5.22, SE=0.221, SD=0.938).

The interaction effect between the two factors was significant ($F(2, 73) = 3.833, p = 0.026$). Fig.4.2 shows that only in the Cooperative condition the explainable negatively influenced the level of trust towards the agents. The specific values per each strategy in the transparent and non-transparent conditions were: Transparent - Cooperative (M=4.90, SE=0.334, SD=1.204), Individualistic (M=4.89, SE=0.224, SD=0.925), Tit for Tat (M=5.51, SE=0.362, SD=1.086) Non-Transparent - Cooperative (M=5.98, SE=0.246, SD=0.602), Individualistic (M=3.81, SE=0.291, SD=1.411), Tit for Tat (M=4.95, SE=0.239, SD=0.711)

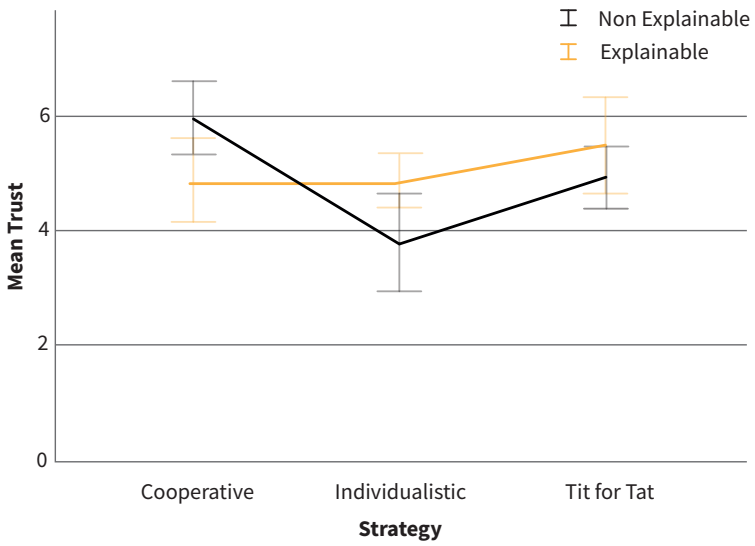

Figure 4.2: Interaction effect between strategy and explainability in trust.

Multi-component Group Identification.

The Group Identification, did not reveal main effect of single factors of explainability and strategy ($F(1, 73) = 2.674; F(3, 73) = 2.360, p = 0.106, p = 0.078$). However, the interaction between the two factors was significant ($F(2, 73) = 4.320, p = 0.017$). The specific values per each strategy in the transparent and non-transparent conditions: Transparent - Cooperative (M=4.15, SE=0.500, SD=1.801), Individualistic (M=5.06, SE=0.336, SD=1.387), Tit for Tat (M=5.27, SE=0.427, SD=1.282). Non-Transparent - Cooperative (M=5.19, SE=0.394, SD=0.965), Individualistic (M=3.32, SE=0.359, SD=1.292), Tit for Tat (M=3.98, SE=0.559, SD=1.676).

As we can notice from the Fig.4.3, explainability and strategy influenced the perception of Group Identification in the opposite direction among the agents' strategies. In the explainable condition, the agents foster less group identification when they acts Cooperatively. However, explainability had a positive influence in the group identification in the Individualistic and Tit for Tat condition. The One-way ANOVA in Group Identification reveals that the effect of explainability in Cooperative condition was not significant ($F(1, 17) = 1.732, p = 0.206$), the effect of explainability in Individualistic condition was significant ($F(1, 28) = 12.178, p = 0.002$) and the effect of explainability in Tit for Tat condition was not significant ($F(1, 16) = 3.398, p = 0.084$).

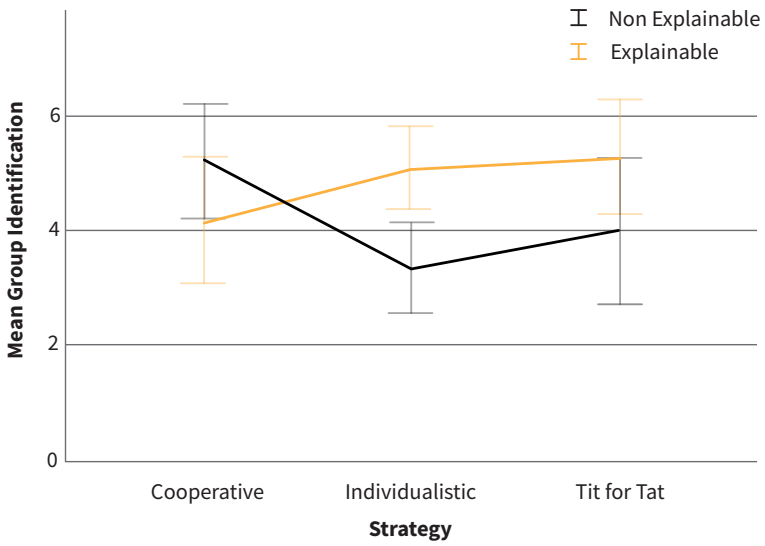

Figure 4.3: Interaction effect between strategy and explainability in group identification.

Goodspeed

The Likeability did not reveal a main effect of explainability ($F(1, 73) = 0.001, p = 0.973$) but informed a main effect of the strategy on the likeability ($F(3, 73) = 3.279, p = 0.026$) Fig.4.5. The interaction between the explainability and strategy was not significant ($F(2, 73) = 0.855, p = 0.429$). Again in this case, the strategy affected the perception of likeability, and no interaction was found regardless of whether or not the agents employ transparent behaviors.

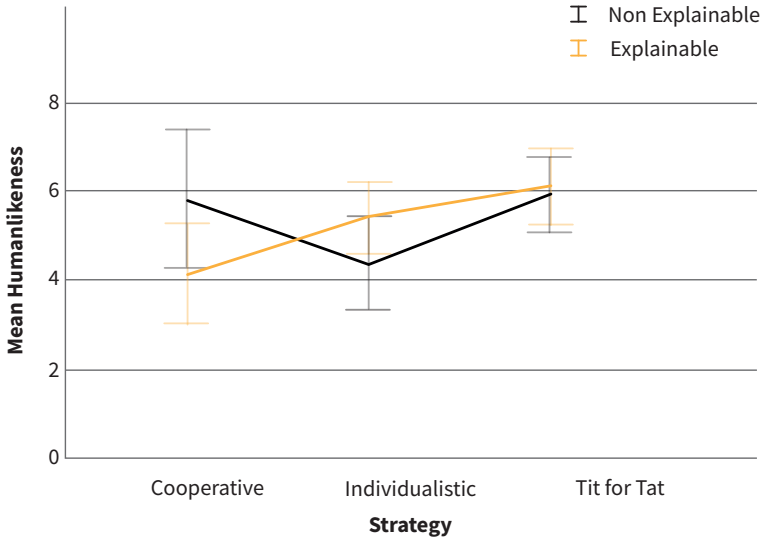

Figure 4.4: Interaction effect between strategy and explainability in humanlikeness.

For the human-likeness dimension, there was no main effect of explainability ($F(1, 73) =$

$0.145, p = 0.704$) and no main effects of the strategy ($F(3, 73) = 2.181, p = 0.098$). However, there was a significant interaction effect between explainability and strategy for the Human-likeness attributed to the agents ($F(2, 73) = 3.585, p = 0.033$).

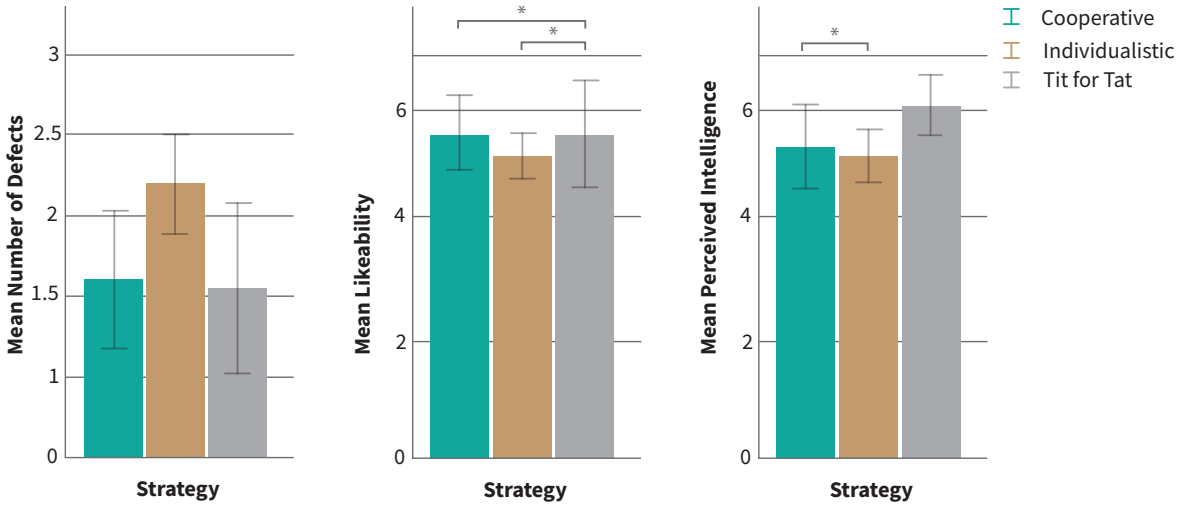

Figure 4.5: Main effect of the strategy on number of defects, likeability and perceived intelligence

In Fig.4.5 we confirmed the trend of a different effect of explainability in the Cooperative condition in respect to the strategy. For the Tit for Tat condition we can notice that both strategy and explainability positively affect the perceived human likeness of the agents.

The Univariate Analysis of Variance of the explainability and strategy for the Perceived Intelligence informed that the main effect of explainability was not significant ($F(1, 73) = 0.652, p = 0.422$), but the main effect of strategy was significant ($F(3, 73) = 5.297, p = 0.002$). The interaction effect between the two fixed factors was not significant ($(2, 73) = 3.632, p = 0.179$). In other words, only the strategy of the agents, regardless of whether or not the agents employ transparent behaviors, affects the perceived intelligence of the agents, in particular for the Tit for Tat strategy as confirmed by several studies about game theory [191][192]. The specific values per each strategy were: Cooperative (M=5.39, SE=0.348, SD=1.518), Individualistic (M=5.23, SE=0.227, SD=1.244), Tit for Tat (M=6.11, SE=0.249, SD=1.054).

4.3 Findings

This chapter explores group interactions involving mixed groups of humans and virtual agents in collaborative game settings. In particular, it is focused on how agents' explainability affects teamwork and the perception of autonomous teammates. Although we have hypothesized that

explainability would positively influence several measures of teamwork, we have also manipulated the strategy of the agents to ascertain if the results would hold similarly when the agents adopted different strategies.

4.3 Human Cooperative Choices

According to **H1**, we expected that the agents' explainability would increase the number of cooperative choices of the human player, which was not confirmed. In fact, we only found a partially significant main effect of the strategy on the number of Cooperative choices, which suggests people cooperated differently according to which strategy the agents adopted. In the post hoc analysis, cooperation towards the Individualistic agents was lower than towards Cooperative and Tit for Tat agents. Additionally, we analyzed the cooperation rate of the agents and we found the Individualistic strategy led the agents to cooperate less compared to the other to strategies, which suggests people might have reciprocated the autonomous agents to a certain extent Fig.4.5. In our experiment, we could not find evidence that explainability affects people's behaviour.

4.3 Explainability Across Strategies

Regarding **H2**, we have hypothesized that trust and group identification would be positively affected by transparent behaviour. On both measures, we found a significant interaction effect of explainability and strategy, which reveals the effect of explainability on trust and group identification was different across the three strategies. In terms of the trust, the post-hoc analysis did not reveal a significant effect of explainability in any of the strategies. However, the trends that are visible in Fig.4.5 suggest this effect was negative for the Cooperative agents and was positive for both the Individualistic and Tit for Tat agents. In the post-hoc analysis for the group identification, we found a significant positive effect of explainability for the Individualistic agents. For the remaining strategies, similar trends are visible in Fig.4.3 suggesting a negative effect for Cooperative agents and a positive effect for Tit for Tat agents. Our hypothesis was only partially validated due to the fact that both group measures showed a positive effect only for two strategies, the Individualistic and Tit for Tat. Later in this section, we discuss the negative effect on the Cooperative strategy.

4.3 Unconditional Cooperators

In **H3**, we have predicted that transparent behaviours would positively affect the likeability and human-likeness of the agents. We only found a significant interaction effect between explainability

and strategy on the perceived human-likeness. In other words, the effect of explainability on the perception of human-likeness was different across the three strategies. Although the post hoc analysis did not reveal a significant effect of explainability in any of the strategies, the trends suggest a negative effect on the Cooperative agents, a positive effect on the Individualistic agents and no effect is suggested for the Tit for Tat agents. In terms of likeability, we found a significant main effect of the strategy with the Individualistic agents being significantly rated as less likeable compared to the Cooperative and Tit for Tat agents This hypothesis was validated in terms of human-likeness for the agents that use a Individualistic strategy.

Our results suggest that adding transparent behaviour to an unconditional cooperator negatively affects the perceptions people have in terms of trust, group identification and human likeness. Although these differences were not statistically significant, the trends are congruent in the same direction. Further investigation is needed to support this claim. In terms of human-likeness, our intuition is that the unconditional cooperator might have revealed to the participants a non-optimal strategy, which a human would probably not do. However, the result for the group measures are counter-intuitive because the non-optimally of this strategy is related to the individual gains and it not clear why the unconditional cooperator negatively affected then perception of the group.

Chapter 5

Explainable Agency by Revealing Suboptimality

Conveying task knowledge through demonstrations alone is challenging. Adding explanations, in particular, contrastive explanations that compare two demonstrations can reduce the complexity of this problem. A natural context to study these problems is within educational scenarios, because the explanations can guide the attention of the learner to specific aspects of the demonstration [193, 194]. Without explanations sub-optimal demonstrations can be easily misconstrued as optimal. Therefore, it is important to understand how to build systems capable of such explanations.

So far, the work that has been done to generate contrastive explanations focused on human inputs, and, to the best of our knowledge, compares alternative plans but does not account for the optimality of the action. Furthermore, few examples in the existing literature of autonomous and explainable robots are tested in a child-robot interaction scenario.

The focus on the human inputs for providing explanations could be explained by the assumption that outside of the educational context, the robot performs optimally with respect to its own understanding of the environment. At the same time, the deployment of explainable robots that are robust enough to work with children is non-trivial.

We address the first challenge by developing an algorithm that returns contrastive explanations comparing optimal and sub-optimal actions. To validate our approach, we deployed our system in a child-robot game scenarios. We compute the robot's explanation using a search-based approach and investigate the effect of the explanation on the child's perceived difficulty of the task, game-play efficiency, and perception of the robot.

As a consequence of our approach we show that it is possible to build a system that informs and explains the reason why its action was sub-optimal. Thanks to our successful deployment in a child-robot educational scenario, our approach is likely robust enough to be applicable to a large range of sequential planning tasks.

Research Questions

This research explores the question of whether a robot that informs and explains the reason why its action was sub-optimal can affect children perceived difficulty of a learning task, game-play efficiency, and perception of the robot.

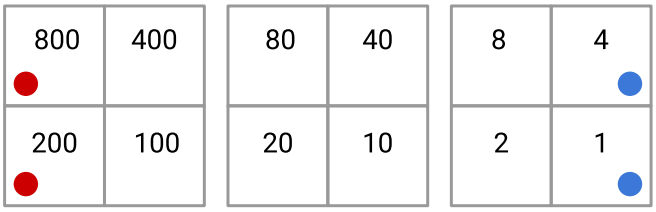

Figure 5.1: Scores associated with each square and the starting position. Note that the scores are not visible to the child.

5.1 Minicomputer Tug of War Game Platform

Game Scenario

As a running example throughout the paper we choose a child-robot interaction scenario based on a two-player zero-sum game called *Minicomputer Tug of War*. The scenario is based on the *Papi's Minicomputer*, a non-verbal language to introduce children to mechanical and mental arithmetic through decimal notation with binary positional rules [1].

The game comprises of three 2×2 square boards. Each of the 12 cells has an associated value. Each player has 2 checkers available, and each checker is worth the value associated with the cell where it stands. One player (the robot) starts with the checkers in the cells 800 and 200 (corresponding to a score of 1000), and tries to minimize its score. The other player (the child) starts with the checkers in the cells 4 and 1 (corresponding to a score of 5), and tries to maximize it's score. The players alternate in moving their checkers; the game ends when the child obtains a higher or equal score of the robot or vice versa. The winner is the player whos turn it is when the game ends. Given the above rules, a state in the game is a configuration of the four checkers. The set of possible states corresponds to all possible configurations of checkers in the 12 cells. At each turn a player is allowed to move one checker to one of the contiguous squares (e.g., it is not possible to move from 1 to 10 or from 8 to 400, while moving from 8 to 40 is legal). The applicable actions for each checker are along the cardinal directions and the diagonals. A player is not allowed to have two of her checkers in one square.

This scenario is useful for our experiment because it has been previously used in the educational context, which means that we can focus on the explainable agency instead of scenario design. As the game scenario is deterministic and adversarial we represent the planning problem as a tree and use minmax for plan generation.

[1] Minicomputer Games, http://stern.buffalostate.edu/CSMPProgram/String, consulted on June 2019

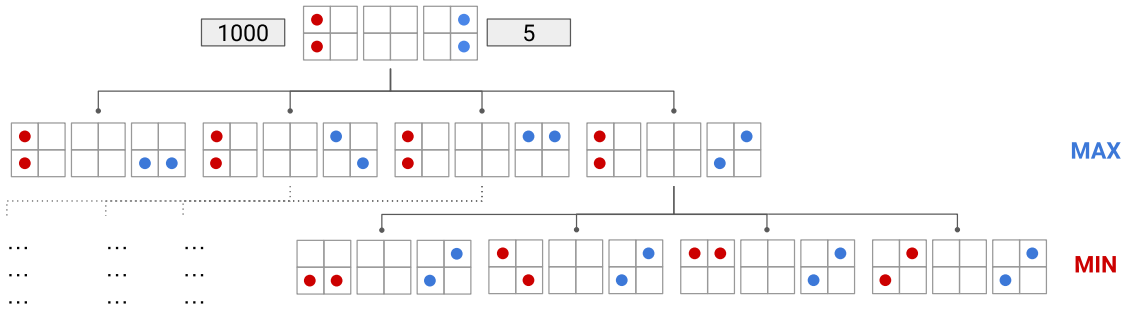

Figure 5.2: A visualization of the tree search. The current state is shown at the top. Each state is expanded for all possible actions following the minmax algorithm.

Search-Based Explanation Generation

We propose a pro-active system to generate explanations to inform the human, about a sub-optimal plan. By pro-active we mean that the system chooses by itself when to give an explanation. The way we achieve this is by planning from the previous state for all possible actions in order to find the optimal plan. To balance the real-time constraints of the system with the high computational cost of planning multiple plans in parallel, we approximate the utility of each plan. We then compare the estimated utility of the current plan with the estimate of the best plan and generate an explanation if the optimal and current action differ. As we have a utility estimate for both plans, the explanation also provides an indicator of the degree of sub-optimality of the performed action.

Consider an adversarial planning problem

$$\mathcal{I} = (P, s_0, \mathcal{A}, U, T), \tag{5.1}$$

where

- P is the set of *players*. In this paper, we consider only two players, denoted as minimizer - MIN (robot) and maximizer - MAX (child opponent).
- s_0 is the *initial state*.
- $\mathcal{A}$ is the finite set of *actions* available to the players. We write $A(s, p)$ to denote the actions available in state s when it is player p's turn to act.
- U is the *utility function*. We write $U(s, p)$ to represent the value/payoff if the game ends in terminal state s. Player MAX wishes to maximize $U(s, p)$, while player MIN wishes to minimize it.

- T denotes the *transition model*. Given a state s and an action a, the transition model returns the subsequent state s' such that $s' = T(s, a)$.

and the set of possible states, $\mathcal{S}$, that is defined by all states reachable from s_0. Further, suppose the agent performs a sub-optimal action a_s in state s.

To generate an explanation and inform the human we start a new planning problem in the previous state by choosing $s_0 = s$, and, for each action a_i compute the maximum utility $v(T(s_0, a_i))$ in the scenario using the minmax formalism

$$v(s) = \begin{cases} U(s, p) & \text{if } s \text{ is terminal;} \\ \max_{a \in \mathcal{A}(s, \text{MAX})} v(T(s, a)) & \text{if } p = \text{MAX acts in } s; \\ \min_{a \in \mathcal{A}(s, \text{MIN})} v(T(s, a)) & \text{if } p = \text{MIN acts in } s. \end{cases}$$

as proposed by Russell and Norvig [195]. Knowing the utility of all actions, we compute the optimal action a^* of the previous state, compare it to the executed one, and provide a contrastive explanation.

In order to compute $v(T(s_0, a_i))$ we use a minmax planner which we limit to a depth bound of $m = 3$ to account for the real-time constraint of our scenario, and then approximate the utility using $U(s, p)$.

To evaluate the above approach, we designed a game scenario in which the robot alternates between choosing optimal and sub-optimal actions. Whenever the robot acts sub-optimally, an explanation is generated. Following the assumption that humans focus on the on abnormal causes to explain events [196], each explanation is introduced as a justification of a mistake.

System Architecture

To enable the robot to play the game and give explanations, we develop a distributed system with three major components: the GAME component, which provides the interface for the user to interact with, the ROBOT component, that controls the embodied platform and the natural-language interface, and the EXPLANATION SYSTEM.

The Explanation System, implemented using the ROS framework, is responsible for planning the agent's actions autonomously and generating the explanation. The system can be decomposed into 5 different modules. The GAME INTERFACE and EXPLANATION MODULE serve as communication modules between the system and the game: the Game Interface receives and updates the players state, while the Explanation Module manages the robot communication and animations, generating human-readable sentences. The PLANNING MANAGER performs a

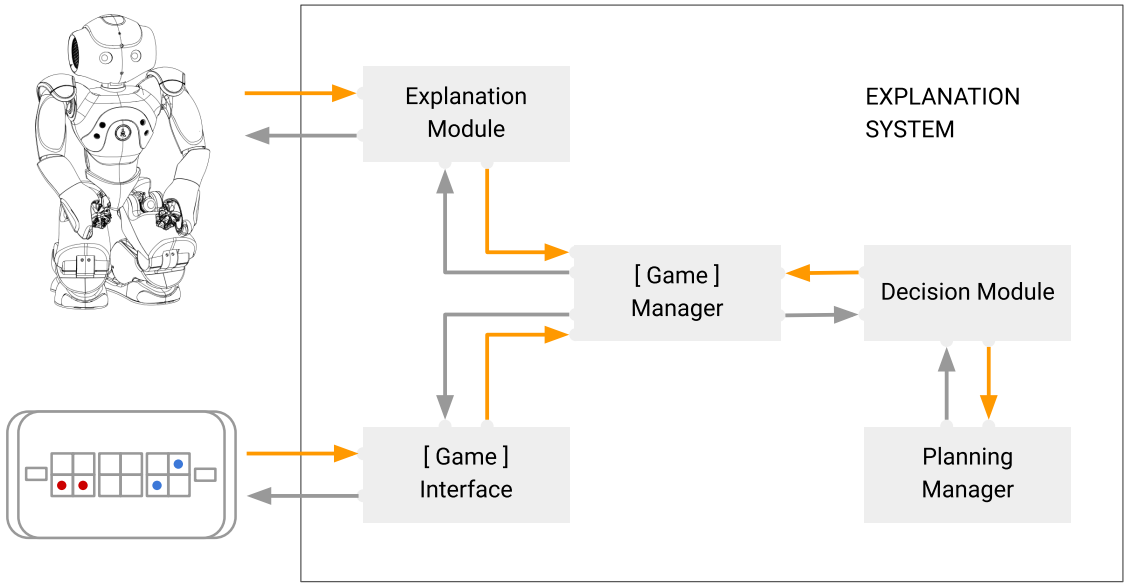

Figure 5.3: A topological overview of the system's architecture.

complete depth-limited exploration of the game tree. The module returns the policy of agent, the action $a \in A(s)$ and the result of each action. The DECISION MODULE selects the agent's policy (optimal or sub-optimal). Finally, the GAME MANAGER links all the above modules: it establishes the starting of the game and the turn-taking events. Moreover, the Game Manager also recalls information from the Decision Module to communicate with the Explanation Module whether the robot action is optimal or not, and consequently, when the explanation is needed.

5.2 Experimental Design

Hypotheses

The hypotheses are that children (H1) will perceive the task as less difficult, (H2) play more efficiently, and (H3) will perceive the robot as more intelligent and animated when it explains its decisions versus when it does not.

Design

To investigate the hypotheses, we asked children to play the *Minicomputer Tug of War* game three times with the robot. To avoid game play loops, we limited children's possible actions to the actions that increase their score. This decision was informed by the result of a pilot study.

Moreover, we assigned the role of maximizer to the child, since subtractions appear later in the curriculum, and introduced the robot as a peer to make mistakes seem natural.

Within the game, we manipulated the robot's explainable agency between participants based on two conditions: (1) the robot does not explain anything (non-explainable), or (2) the robot explains its sub-optimal action in comparison to the optimal contrast case (explainable). To give explanations, we used templates such as:

> "I made a mistake. I moved the ball + [checker] + to + [action] + and I'm going to obtain [score] + points in the next + [number of turns] + turns, but moving the ball [best checker] + to + [best action] + I could have gotten + [score] + points in the next + [number of turns] + turns. Now it is your turn."

We recorded a variety of dependent variables to assess our hypotheses. To measure the perceived difficulty (H1), we asked children to solve six exercises validated by their teachers and related to the abacus system, due to the strong similarity with the game, and compared the scores in a pre- and post-test. We also asked children to report the perceived difficulty on performing the tests. To measure efficiency (H2) we recorded the number of moves until completion of the game, the score obtained after each move, and how often the child won. To measure the perception of the robot (H3) we provided a revised version of the Godspeed questionnaire [197][2].

Finally, we asked five exploratory questions to learn more about children's perception of explainable agency.

To comply with local regulation, the work described has been carried out in accordance with *The Code of Ethics of the World Medical Association* (Declaration of Helsinki) for experiments involving humans; informed consent has been obtained for experimentation with human subjects. The privacy rights of human subjects has been always considered. We informed both the parents and the children about the confidentiality of the data, the voluntary participation and the authorization for sharing the results with the purpose of analysis, research and dissemination.

Participants

Participants were 33 children from a school that integrates the *Papi's Minicomputer* (abacus system) in their curriculum. All participants attended 2nd grade and were randomly assigned to one of the two conditions. One child was excluded, because of technical difficulties, and we analyzed the data from the remaining 32 (age $M = 7.03; SD = .18$, gender [non-explainable: 10 Male, 6 Female, explainable: 8 Male, 9 Female]).

[2]The questionnaires are shared in the supplementary material: Cloud Folder

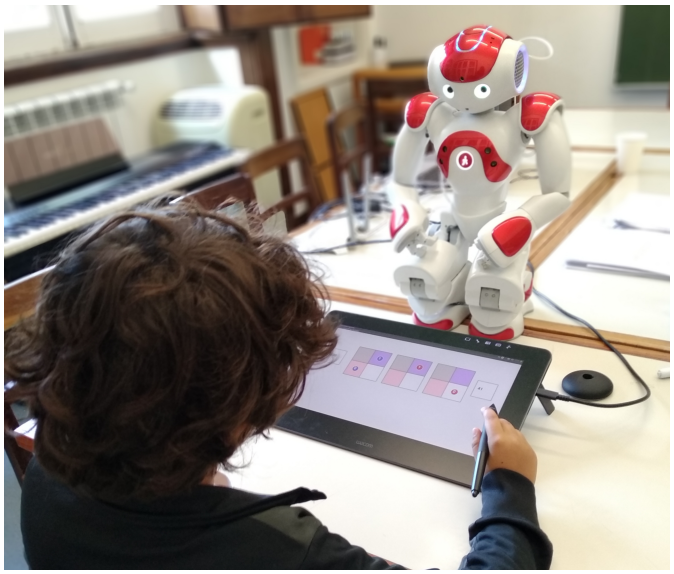

Figure 5.4: Deployment of our explanation generation system in an educational scenario.

Materials

Children played the game and interacted with the proposed system (Fig. 5.1). NAO sat on the table in a crouching position opposite a Wacom Cintiq Pro 13 Tablet with pen. The trial took place in a separate room of the school, and was conducted in the local language (portuguese).

Procedure

The experiment began by randomly assigning a child to one of two conditions. Children entered the room, and were asked to sit in front of the robot (Fig. 5.4). The researcher explained that before talking to the robot they were going to answer a few simple questions (pre-test). Once the pre-test was done the robot introduced itself and asked for the child's name. Once the child answered, the robot asked if the child knew *Papi's Minicomputer* (abacus system), and if they ever used it to play *Minicomputer Tug of War* (the game) and proceeded explaining the game rules. Once the robot finished, the researcher made sure that the child understood the instructions and the game began. The child and robot took turns playing three games; after the games the researcher told the child that the game was over, and they had to answer more questions (post-test). Once the child completed the questionnaires the researcher asked if they had any questions and thanked them for their help. The sessions were individual and took approximately 20 minutes to complete.

Results

To assess perceived difficulty (H1) we scored the pre- and post-test assigning 1 point for correct answers and 0.5 points when they mirrored the abacus system. We then summed up the values to obtain a final score. A Wilcoxon's t-test between pre- ($M = 4.18; SE = .21$) and post-test ($M = 4.82; SE = .13$) revealed a significant difference ($Z = -2.6; p = .008$) for the explainable group (Fig. 5.5). No difference was found for the non-explainable group ($p \geq .05$). Here higher scores indicate that the task was perceived as easier.

Regarding efficiency (H2), a multivariate analysis of variance on the the number of moves until completion of the game, the score obtained after each move, and how often the child won, was not significant (*Pillai Trace* $= .12068, F(3, 30) = 1.2809, p = .3002$). There was also no main effect for the explainable group in the score obtained after each of the three initial turns of the game play (*Pillai Trace* $= .24946, F(9, 27) = 0.70169, p = .7002$; nor on the number of wins ($F(1, 30) = 0.039, p = 0.845$). We then compared the proportion of games won, by condition. This analysis yield no significant results ($U = 114.5; W = 234.5; p = .621$).

To investigate the perception of the robot (H3) we firstly analyzed the reliability of the Godspeed. The reliability analysis revealed low internal consistency among items (GODSPEED: $\alpha = 0.61$), for perceived intelligence ($\alpha = 0.53$), and for animacy ($\alpha = 0.36$). No differences were found in both groups, for the different dimensions of the questionnaire (all $p \geq .05$). We calculated the correlation between perceived intelligence and the proportion of wins ($\rho = -.33; p$=*.068*). Although it is only marginally significant it shows that children that win more games, perceive the robot as less intelligent than children who win less games.

Looking at the exploratory questions. We divided the children's answers to the open ended questions in four categories depending on what they reported to be helpful. In the explainable condition, the majority of the answers affirmed that the robot's explanations during the game play and its way of playing were supportive, while the 15% of the children stated that they were aided by the robot's explanation of the rules (explainable: robot's speech = 50%, robot's way of playing = 35%, robot's explanation of the game = 14%; *non-explainable condition*: robot's speech = 10%, robot's way of playing = 50%, robot's explanation of the game = 30%, robot's gesture = 10%). The 10% of the children in the *non-explainable condition* reported that the robot gaze was useful.

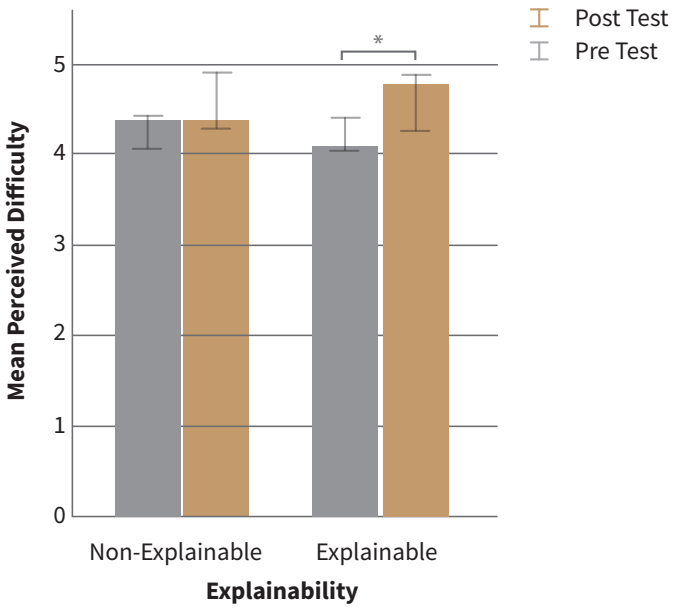

Figure 5.5: Perceived difficulty of the pre- and post-test by condition: Explainable, Non-Explainable

5.3 Findings

Throughout this research, we have deployed our approach and demonstrated its applicability to a real-world scenario. This shows that our system is robust enough for interaction with children in the wild, even though is autonomous.

5.3 Task Difficulty

In (H1) we predicted that the robot's explanations would affect the children's perception of the task difficulty. Indeed we found a significant positive effect of explainable agency on the perceived difficulty of the pre- and post-test. Hence, we can follow that explainable agency has a positive impact in those scenarios.

Our results are in line with the idea of self-efficacy by Bandura [198]. Self-efficacy relates to people's beliefs about their own capabilities and it is intimately connected to agency (e.g. Pastorelli et al. [199]). The information provided by the robot in the explainable condition may have served as feedback about children's efficacy in using the abacus system which in turn made them feel more confident about their capabilities (higher self-efficacy).

5.3 Children's Efficiency

According to (H2), we expected the robot's explanations to improve the children's efficiency during game-play. However, the data did not confirm this hypothesis. This lack of significance may be explained by the limited set of actions available to the child. This may have influenced the variability of the data collected, and consequently the observable differences between the two conditions. Further investigation is needed to support this claim.

5.3 Robot's Intelligence

Regarding (H3), we have hypothesized that the robot's explanation influences the child's perception of the robot's intelligence and animacy. Overall, we did not find significant effect of the explainable agency. We assume that the effect of explainable agency is less strong than other social factors, such as gestures. Future work should consider these effects for example by exploiting multimodal non-verbal behaviors to support a more clear explanation.

Nevertheless, the answers to the open question about the robot's explanation provide interesting cues. In the explainable condition, the children reported that the robot helped them by showing the best action or mentioning possible alternative actions (e.g., "Showed me the best I could do and how to play",“The robot told me how it could get more points", “The robot told me that if it moved differently it would have gotten more points"). What we considered explainable agency - the robot's speech during the game - might have been perceived differently by the children. Some children appeared to consider that the robot was making mistakes to help them play (e.g. “The robot helped by playing badly", “Doing mistakes"), which seems to take the focus away from the speech and into its actions. Another tendency was the children referring to the robot explaining the rules of the game as being helpful and not referring to the robot's speech during the game.

Chapter 6

Learning from Explanations as Inverse Planning

As robots and other autonomous agents enter our homes, hospitals, schools, and workplaces, it is critical to find ways to adapt their behaviour to tasks that occur unexpectedly, thus *learning* through natural, real-time interactions with the environment and its inhabitants [14]. There exist several learning strategies used by both humans and agents to learn a new task.

An approach to learning from other agents is imitation. Enabling machines to learn a desired behavior by imitating an expert's behavior has been proven to be a powerful tool to speed up the learning process [119]. This approach is inspired by human *imitation learning* (IL) processes, and is also known as *learning from demonstration* (LfD) [18, 125], *programming by demonstration* (PbD) [200], and *teaching by showing* [201].

To exemplify the intents behind the experts' demonstrations and direct the learner towards crucial aspects of the task, humans often substitute or complement rewards and demonstrations with other teaching signals, such as explanations. The act of explaining can be thought of as a mean to transfer knowledge between an explainer, i.e., someone who is in possession of explanatory information, and an explainee, i.e., someone or a group of people who is thought not to possess it already [27, 28]. This process has been identified as the *social process* of the explanation [29]. In a continuous interaction between the explainer and her counterpart, the main goal of the explainer is to provide enough information to the explainee so that they can understand the causes of some fact or event. This process contemplates the active role of the explainee, which can ask for explanations by querying the explainer. In addition to the *social process*, explanation has been described also as a *cognitive process* and a *product* [9, 30]. The *cognitive process* concerns with abductive inference, a form of logical inference that starting from the observation or set of observations seeks for the simplest and more likely conclusion, i.e., explanatory hypotheses [31, 32].

In summary, explanations describe how and why something works the way it does, allowing humans to solve ambiguities in their current knowledge state [202, 203] and evaluate observed actions.

In the context of intelligent agents, explanations could be a valuable way to concisely describe a task or extrapolate useful information from a set of demonstrations, thus decrease the number of examples needed to replicate a behavior and generalize a certain knowledge to unseen situations.

There exist two compelling lines of research concerned respectively with how to incorporate the knowledge of an expert, and how to summarize the behavior of an expert.

So far, the work that has been done shows that inverse reinforcement learning (IRL) algorithms are helpful to integrate various types of previous knowledge [204, 205, 130, 206]. The expert's knowledge is encoded in a set of demonstrations, each demonstration comprises samples

exemplifying the behavior of the expert, e.g., the action selected in a specific state. To reduce the number of samples needed to learn the observed task, inverse reinforcement learners can query the demonstrator about specific states [136], rank demonstrations to extrapolate the underlying intent of the best demonstration [207], or learn progressively more challenging source tasks [208].

Moreover, the possibility to integrate statistical methods to estimate the parameters of an assumed probability distribution, given some demonstrations, allows for further improvements in the performance of IRL agents [209].

Solutions for explaining the decisions of sequential decision making agents include: techniques to answer questions such as *"Why has this recommendation been made?"* by populating generic templates with domain-specific information from the task [210, 74], approaches to map action queries such as *"When do/will you <action>?*" into policy explanations by inspecting the states in which the input action is the most likely one [58, 56, 71], and methods to generate counterfactual explanations of behaviour based on the causal relationships between variables of interest [211, 11].

Human evaluations of agents' explanations generally take the form of a user study and examine the knowledge gained through task prediction performance, i.e., the explainee, recipient of the explanations, would be able to provide a better prediction if explanations successfully made the model intelligible. Comparatively, to the best of our knowledge, the evaluation of agents' explanations as teaching signals in learning and teaching scenarios is hardly explored.

This work aims at understanding whether reasoning upon explanations of an expert would make a learning agent more efficient, and validating how this learning approach would be valuable in teaching scenarios involving humans.

Similarly to Juozapaitis et al. [74], our work provides information about the positive or negative valence of an action in a certain state by comparing the q-values of two possible actions. Thanks to this comparison our system builds explanations similar to the ones generated by causal approaches, thus replying to *why* and *why not* questions. In addition, our approach accounts for the goodness of the state and includes a parameter to indicate how trustworthy the explanations are.

Differently from previous works [156, 11], we evaluate explanation against other types of teaching signals, i.e., reward and demonstration, controlling for the situation and the position of the learner with respect to the goal.

First, we introduce and formalize a method to integrate explanations into maximum likelihood inverse reinforcement learning. Second, we computationally evaluate our method on three

navigational scenarios using three different types of teaching signals, i.e., reward, demonstration, explanation. Results indicate that explanations lead to better performance in all scenarios. Finally, we conduct a user study using the implemented teaching signals and evaluate participants' preferences in four different situations each constituted by a set of eight positions of the learner with respect to the goal. Results show that explanations are preferred when the learner is far from the goal.

Research Questions

This work explores the following research questions:

- Can agents' learn more efficiently with explanations?
- Do humans generally prefer explanations to teach each other how to solve a task?
- Does the type of teaching signals preferred by humans depend by the contextual situation?

Our research effort takes inspiration from human social learning mechanisms to focus on situations in which an expert guides a learner through explanations. The proposed approach incorporates explanations into maximum likelihood inverse reinforcement learning. We computationally evaluate explanations against other teaching signals (reward, demonstration and explanation) in three navigational scenarios. The generated explanations are also evaluated in a user study with 150 participants. The user study investigates participants' preferences between the different types of teaching signals and the impact of contextual situations, i.e., distance from the task's goal, on their preferences. Our simulations' results show that explanations lead to better performance compared to reward and demonstration signals, and that explanations are preferred by human teachers in situations where the goal is far from the learner.

6.1 Background

Markov Decision Processes

Sequential decision problems incorporate utilities, uncertainty, and sensing, and include search and planning problems as special cases [212]. Sequential decision problems can be formalized as Markov Decision Processes (MDPs) [213]. In MDPs, the agent's utility depends on a sequence of decisions. A MDP encodes the sequential decision making as a tuple $(\mathcal{S}, \mathcal{A}, \{P_a\}, r, \gamma, \mu)$ where $\mathcal{S}$ represents the state-space, $\mathcal{A}$ the action-space, $\{P_a, a \in \mathcal{A}\}$ denotes the set of transition probabilities defining the dynamics of the MDP, i.e.,

$$P_a\left(s' \mid s\right) \triangleq P_r\left(S_{t+1} = s' \mid S_t = s, A_t = a\right) \tag{6.1}$$

$r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\gamma \in [0, 1)$ is the discount factor, and $\mu : \mathcal{S} \to \mathbb{R}_{\geq 0}$ is the initial state distribution. The goal of the decision-maker is to determine the series of actions that maximize the agent's total discounted reward, i.e. $TDR = E\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right]$ where R_{t+1} is the random reward received by the agent at time step $t + 1$ as a consequence of performing some action A_t in state S_t. The principle used to select a series of actions, i.e., the policy, is a mapping $\pi : S \times \mathcal{A} \to [0, 1]$. A policy π can be described as the probability of choosing a certain action in a given state $P_a\left[A_t = a \mid S_t = s\right] = \pi(s, a)$.

Value Function and Optimality

The value $v^\pi(s)$ of a policy π can be defined as:

$$v^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r\left(S_t, A_t\right) \mid S_0 = s\right].$$

where v^π is the sum of the discounted rewards for all time steps, and γ is the discount factor assigning the importance to the rewards obtained over time. Whereas the reward signal r specifies what is good in an immediate sense, the *value function* $v^\pi(s)$ specifies what is good in the long run [132].

The optimal value function verifies the recursive relation:

$$v^{\pi^*}(s) = \max_{a \in \mathcal{A}}\left[r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P_a\left(s' \mid s\right) v^{\pi^*}\left(s'\right)\right]$$

Conversely, the optimal q-function, or action-value function, is the value that the decision-maker expects to collect starting from state s, taking action a, following policy π. The optimal q-function is defined as:

$$q^*(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P_a\left(s' \mid s\right) v^{\pi^*}\left(s'\right)$$

The q-function summarizes all the relevant information the agent has to know by providing a ranking for the actions according to how useful they are for a particular goal that the agent has. By using this function, at each state the agent can search for the action that has the maximum q^* value. Both the optimal policy π^* and v^{π^*} can be computed from q^* as:

$$\pi^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}} q^*(s, a)$$

$$v^*(s) = \max_{a \in \mathcal{A}} q^*(s, a)$$

6.2 Learning from Explanations

We present a maximum likelihood inverse reinforcement learning approach to allow agents to learn from explanation. We formalize the problem of *learning from explanations* (LfE) as an inverse reinforcement learning problem and use ideas from optimal control such as the incompletely-known *Markov Decision Process*, and *value function* (detailed in section 6.1 and section 6.1 respectively).

We assume that the reward function to be learned, henceforth denoted as r^*, can be represented as a linear combination of features. To estimate the value of the linear combination of features describing the expert's reward function we use *maximum likelihood estimation* (MLE).

Given that the reward function f and the dynamics of the system M are known linear mappings and assuming the data is large enough, finding θ involves solving a system of linear equations. Maximum likelihood estimation has been used in previous work in inverse reinforcement learning, more details on this can be found in chapter 2.

Within this framework we define three different teaching signals: reward, demonstration, explanation. A reward signal emulates the reinforcement learning approach and includes information about the state, the action and the reward associated with the state-action pair. A demonstration signal mimics the learning from demonstration approach and constitutes of information about the state and the action. Finally, a explanation signal gives information about the state, the action, the contrastive action, the next state and the goodness of that state.

6.2 Learning a task

Throughout this document, we consider a learner who knows a rewardless MDP, i.e., $(\mathcal{S}, \mathcal{A}, \{P_a\}, \gamma)$, and must learn a task description in the form of a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. We assume that the reward to be learned, henceforth denoted as r^*, can be represented as a linear combination of features, i.e., given a set of K features $\phi_1, \ldots, \phi_K$ with $\phi_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for $k = 1, \ldots, K$,

$$r^*(s, a) = \sum_{k=1}^{K} \phi_k(s, a) w_k^* = \phi^\top(s, a) w^*$$

for some weight vector $w*$. We denote by v_w^*, π_w^*, and q_w^* the optimal value function, policy, and q-function for the MDP $\left(\mathcal{S}, \mathcal{A}, \{P_a\}, \phi^\top w, \gamma\right)$. The goal of the learner is, therefore, to recover some weight vector w^*.

6.2 Learning a task from rewards

The first approach to learning the task is given samples of the reward. A sample consists of a triplet (s, a, r), where r is a reward observed upon performing a in state s. Specifically, we assume that the sample rewards r correspond to independent observations of $r^*(s, a)$ corrupted by zero-mean Gaussian noise with known precision η, so that:

$$\Pr\left(s, a, r \mid r^* = \phi^\top w\right) = \text{Normal}\left(r - \phi^\top(s, a)w; 0, \eta\right) \tag{6.2}$$

The maximum likelihood estimate for w^* thus comes:

$$\begin{aligned}\hat{w}^* &= \text{argmax}_{w\in\mathbb{R}^K} \prod_{n=1}^{N} \text{Normal}\left(r_n - \phi^\top(s_n, a_n)\, w; 0, \eta\right) \\ &= \text{argmax}_{w\in\mathbb{R}^K} \sum_{n=1}^{N} \log \text{Normal}\left(r_n - \phi^\top(s_n, a_n)\, w; 0, \eta\right) \\ &= \text{argmin}_{w\in\mathbb{R}^K} \sum_{n=1}^{N} \left(r_n - \phi^\top(s_n, a_n)\, w\right)^2\end{aligned} \tag{6.3}$$

Then, given a set of samples $\{(s_n, a_n, r_n), n = 1, \ldots, N\}$, we want to compute a weight vector w to minimize the loss:

$$L(w) = \sum_{n=1}^{N} \left(r_n - \phi^\top(s_n, a_n)\, w\right)^2 \tag{6.4}$$

Using standard stochastic gradient descent we get the online update:

$$w_{n+1} = w_n + \alpha_n \phi(s_n, a_n) \left(r_n - \phi^\top(s_n, a_n)\, w_n\right) \tag{6.5}$$

6.2 Learning a task from demonstrations

The second approach we consider is to recover the task from sample demonstrations. We consider a demonstration as a pair (s, a), indicating that the optimal action in state s is a. We assume that the demonstrations are independent and subject to noise, such that

$$\Pr\left(s, a \mid r^* = \phi^\top w\right) = \sigma_w(s, a; \eta) \triangleq \frac{\exp\left\{\eta Q^*_w(s, a)\right\}}{\sum_{a'\in\mathcal{A}} \exp\left\{\eta Q^*_w(s, a')\right\}}, \tag{6.6}$$

where η is a *confidence parameter*, indicating how trustworthy the demonstrations are. Smaller η

allows for more imprecise demonstrations, while larger η requires more precise demonstrations. The expert draws actions from a Boltzmann distribution (softmax) over the learned q-values. The maximum likelihood estimate for w^* thus comes

$$\begin{aligned}\hat{w}^* &= \sum_{n=1}^{N} \log \sigma_w \left(x_n, a_n; \eta\right) \\ &= \operatorname{argmin} \sum_{n=1}^{N} \left(\log \sum_{a' \in \mathcal{A}} \exp\left\{\eta Q_w^* \left(s_n, a'\right)\right\} - \eta Q_w^* \left(s_n, a_n\right) \right).\end{aligned} \tag{6.7}$$

Then, given a set of demonstrations $\{(s_n, a_n), n = 1, \ldots, N\}$, we want to compute w to minimize the loss

$$L(w) = \sum_{n=1}^{N} \left(\log \sum_{a' \in \mathcal{A}} \exp\left\{\eta Q_w^* \left(s_n, a'\right)\right\} - \eta Q_w^* \left(s_n, a_n\right) \right). \tag{6.8}$$

For a single sample (s_n, a_n),

$$\frac{\partial L(w)}{\partial Q_w^* \left(s_n, a\right)} = \eta \left(\sigma_w \left(s_n, a; \eta\right) - \delta_{a, a_n}\right) \tag{6.9}$$

where δ_{a,a_n} is the Kronecker delta function. On the other hand, let

$$P_{\pi_w^*} \left(s' \mid s\right) = \sum_{a \in \mathcal{A}} \pi_w^*(a \mid s) P_a \left(s' \mid s\right) \tag{6.10}$$

and let $\Phi_{\pi_w^*}$ denote $|\mathcal{S}| \times K$ with s, k element given by

$$\left[\Phi_{\pi_w^*}\right]_{s,k} = \phi_{\pi_w^*, k}(s) \stackrel{\Delta}{=} \sum_{a \in \mathcal{A}} \pi_w^*(a \mid s) \phi_k(s, a). \tag{6.11}$$

We have that

$$\frac{\partial Q_w^* \left(s_n, a\right)}{\partial w_k} = \phi_k \left(s_n, a\right) + \gamma P_a \left(s_n\right) \left(I - \gamma P_{\pi_w^*}\right)^{-1} \Phi_{\pi_w^*, k}, \tag{6.12}$$

where $P_a(s)$ is (row) vector corresponding to sth row of P_a. We ignored the dependence of π_w^* on w. This finally yields

$$\nabla_w L(w) = \sum_{a \in \mathcal{A}} \eta \left(\phi \left(s_n, a\right) + \gamma \left(P_a \left(s_n\right) \left(I - \gamma P_{\pi_w^*}\right)^{-1} \Phi_{\pi_w^*} \right)^{\top} \right) \left(\sigma_w \left(s_n, a\right) - \delta_{a, a_n}\right) \tag{6.13}$$

and we get the update

$$w_{n+1} = w_n + \alpha_n \sum_{a \in \mathcal{A}} \eta \left(\phi\left(s_n, a\right) + \gamma \left(P_a\left(s_n\right) \left(I - \gamma P_{\pi_w^*}\right)^{-1} \Phi_{\pi_w^*} \right)^\top \right) \left(\sigma_w\left(s_n, a\right) - \delta_{a,a_n}\right) \quad (6.14)$$

6.2 Learning a task from explanations

We finally consider learning a task from explanations. We consider an explanation a tuple (s, a, b, s', v) with $s, s' \in \mathcal{S}$, $a, b \in \mathcal{A}$, and $v \in \{-1, +1\}$ determines the positive or negative valence of the explanation (good/bad). The natural language explanation can be written with following semantics:

- In state s action a is *better* than action b because it will eventually lead you through state s' and that is *good*.
- In state s action a is *worse* than action b because it will eventually lead you through state s' and that is *bad*.

We consider that, given the target reward r,

$$r_{\pi^*}(s) = \sum_{a \in \mathcal{A}} \pi^*(a \mid s) r(s, a) \quad (6.15)$$

- A state s is *good* in the above sense, if $r_{\pi^*}(s) > 0$ for the optimal policy π_r^* given that reward.
- A state s is *bad* in the above sense, if $r_{\pi^*}(s) < 0$ for the optimal policy π_r^* given that reward.
- An action a is *better* than b in state s and leading to state s' if the following two conditions are cumulatively met:
  - $Q^*(s, a) > Q^*(s, b)$ (a is better than b)
    - The transition probabilities by first taking action a and then following π^* lead to larger transition probability from s to s' than those same probabilities taking action b.
- An action a is *worse* than an action b in state s and leading to state s' if the following two conditions are cumulatively met:
  - $Q^*(s, a) < Q^*(s, a')$ (a is worse than b)

* The transition probabilities by taking action a in s and following π^* elsewhere lead to larger transition probability from s to s' than those same probabilities taking action b.

- Finally, we consider that an action a in state s leads to state b if the policy $\hat{\pi}$ defined as

$$\hat{\pi}\left(s'\right) = \begin{cases} a & \text{if } s' = s \\ \pi^*\left(s'\right) & \text{otherwise} \end{cases} \tag{6.16}$$

is such that

$$P_a^{\infty}(s' \mid s) \stackrel{\Delta}{=} (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P_{\hat{\pi}}^t(s' \mid s) > 0. \tag{6.17}$$

Let

$$\sigma(z;\eta) = \frac{1}{1+\exp\{-\eta z\}} \tag{6.18}$$

where η is *confidence parameter*, indicating how trustworthy the explanations are. We consider

$$\Pr\left(s,a,s',b,v \mid r^* = \phi^\top w\right) = \sigma\left(v r_{\pi_w^*}(s');\eta\right)\sigma\left(v\left(Q_w^*(s,a) - Q_w^*(s,b)\right);\eta\right)P_a^{\infty}(s' \mid s) \tag{6.19}$$

assuming, as before, that the explanations are independent and potentially *noisy*. Then,

$$\begin{aligned} -\log\Pr\left(s,a,s',b,v \mid r^* = \phi^\top w\right) = &\log\left(1+\exp\left(-\eta v r_{\pi_w^*}(s')\right)\right) \\ &+ \log\left(1+\exp\left(-\eta v\left(Q_w^*(s,a) - Q_w^*(s,b)\right)\right)\right) \\ &- \log\left(P_a^{\infty}(s' \mid s)\right) \end{aligned} \tag{6.20}$$

Then, given a set of explanations, $\{(s_n, a_n, b_n, s'_n, v_n), n = 1, \ldots, N\}$, we get the loss function

$$
\begin{aligned}
L(w) = & \sum_{n=1}^{N} \log\left(1 + \exp\left(-\eta v r_{\pi_w^*}\left(s'_n\right)\right)\right) \\
& + \log\left(1 + \exp\left(-\eta v \left(Q_w^*(s,a) - Q_w^*(s,b)\right)\right)\right) \\
& - \log\left(P_a^{\infty}(s' \mid s)\right)
\end{aligned}
\tag{6.21}
$$

which can be optimized online using standard stochastic gradient descent. We have

$$
\nabla_w \log\left(1 + \exp\left(-\eta v r_{\pi_w^*}(s')\right)\right) = \eta v \phi_{\pi_w^*}(s')\left(\sigma\left(v r_{\pi_w^*}(s'); \eta\right) - 1\right) \tag{6.22}
$$

where we again disregard the dependence of π_w^* on w. Similarly,

$$
\begin{aligned}
\nabla_w \log(1 + \exp(-\eta v(Q_w^*(s,a) - Q_w^*(s,b)))) \quad & = \eta v(\phi(s,a) - \phi(s,b) \\
+ \gamma((P_a(s) - P_b(s))(I - \gamma P_{\pi_w^*})^{-1}\Phi_{\pi_w^*})^{\top})(\sigma(v(Q_w^*(s,a) - Q_w^*(s,b)); \eta) - 1)w &
\end{aligned}
\tag{6.23}
$$

Finally, we have that,

$$
P_a^{\infty} = (I - \gamma P_{\hat{\pi}})^{-1} \tag{6.24}
$$

and the computation of $\frac{\partial P_a^{\infty}}{\partial w_k}$ is far from trivial, since the dependence of $\hat{\pi}$ on w is highly nonlinear and, in general, non-differentiable. To make the computation feasible, we instead consider a smooth approximation to $\hat{\pi}$, whereby

$$
\hat{\pi}\left(a' \mid s'\right) \approx \begin{cases} 1.0 & \text{if } s = s' \text{ and } a = a' \\ \sigma_w\left(a' \mid s'\right) & \text{otherwise} \end{cases} \tag{6.25}
$$

with σ_w defined in 6.2.2 Then,

$$
\begin{aligned}
\frac{\partial P_a^{\infty}}{\partial w_k} &= \gamma\left(I - \gamma P_{\hat{\pi}}\right)^{-1} \frac{\partial P_{\hat{\pi}}}{\partial w_k}\left(I - \gamma P_{\hat{\pi}}\right)^{-1} \\
&= \gamma P_a^{\infty} \frac{\partial P_{\hat{\pi}}}{\partial w_k} P_a^{\infty}
\end{aligned}
\tag{6.26}
$$

with

$$
P_{\hat{\pi}}(s' \mid s) = \sum_{a \in \mathcal{A}} \hat{\pi}(a \mid s) P_a(s' \mid s). \tag{6.27}
$$

This yields

$$\begin{aligned}\nabla_w P_{\hat{\pi}}(s' \mid s) &= \sum_{a \in \mathcal{A}} \nabla_w \hat{\pi}(a \mid s) P_a(s' \mid s) \\ &= \sum_{a \in \mathcal{A}} \frac{\hat{\pi}(a \mid s)}{\hat{\pi}(a \mid s)} \nabla_w \hat{\pi}(a \mid s) P_a(s' \mid s) \\ &= \sum_{a \in \mathcal{A}} \hat{\pi}(a \mid s) \nabla_w \log \hat{\pi}(a \mid s) P_a(s' \mid s).\end{aligned} \tag{6.28}$$

Using the results from learning from demonstrations method,

$$\begin{aligned}\frac{\partial P_{\hat{\pi}}(s' \mid s)}{\partial w_k} = \sum_{a \in \mathcal{A}} \hat{\pi}(a \mid s) \sum_{a' \in \mathcal{A}} \eta(\phi_k(s, a') + \gamma(P_{a'}(s) P_a^{\infty} \phi_{\pi^*_w, k})^{\top})(\sigma_w(s, a') \\ - \delta_{a,a'}) P_a(s' \mid s)\end{aligned} \tag{6.29}$$

Finally, putting everything together,

$$\frac{\partial \log P_a^{\infty}(s' \mid s)}{\partial w_k} = \frac{\gamma}{P_a^{\infty}(s' \mid s)} \left(P_a^{\infty} \frac{\partial P_{\hat{\pi}}}{\partial w_k} P_a^{\infty} \right)_{s,s'} \tag{6.30}$$

6.3 Experiments

6.3 Simulation Experiment

We evaluate the learning from explanations (LfE) framework to determine if it leads to better performance compared to other types of learning approaches, i.e., learning from rewards and demonstrations. The effectiveness of the learned policy is evaluated based on a agent's capability of using the learned reward in the original problem. We refer to this capability as the performance of the agent. Our hypothesis **(H1)** is that agents will learn more efficiently from explanations than from both rewards and demonstrations.

Methodology

We consider the environments depicted in Figure 6.4. The agent operates in a grid, and is able to move in four directions, i.e., *up*, *down*, *left*, *right*, or stay in place *stay*. Each action moves the agent deterministically to an adjacent cell, factoring in obstacles. The agent is rewarded for navigating from an initial random state A to the goal B, which is one of the colored cells, in the

most efficient way. In the simulation, the movement actions succeed with probability 0.8 and fail with probability 0.2. The reward is a linear combination of features, each corresponding to the indicator for one of the states in orange. Each environment has a different configuration. Environment 6.1 is a 5×5 grid with 19 possible states and four objects. Environment 6.2 is a 3×5 grid with 12 states and three objects. Environment 6.3 is a 6×5 grid with 15 states and three objects. The combination of corridors and objects simulates ambiguous situations in which explanations might help in choosing the best action depending on the state of the agent and the reward associated with each of the objects.

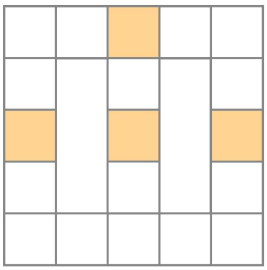

Figure 6.1: Env 1 - Grid-world Environment Consisting of 19 States and Four Objects

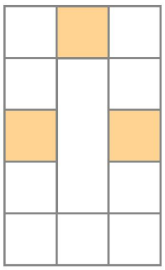

Figure 6.2: Env 2 - Grid-world Environment Consisting of 12 States and Three Objects

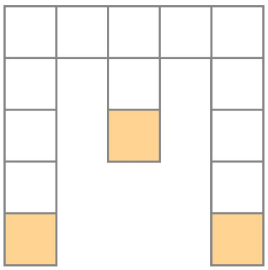

Figure 6.3: Env 3 - Grid-world Environment Consisting of 15 States and Three Objects

Figure 6.4: Navigational environments used for the computational evaluation of the learning from explanation (LfE) framework.

We designed an experiment with three conditions: (1) learning from rewards (LfR), (2) learning from demonstrations (LfD), (3) learning from explanations (LfE). At the beginning, we generate a random reward as a convex combination of features, and compare the performance of an expert, a LfR updater, a LfD updater and a LfE updater in randomly selected sample rewards, demos and explanations.

We start by defining the samplers. The reward sampler selects a random state s and a random action a from the given MDP M and provides the corresponding noisy reward r as a linear combination of features from a standard normal distribution. The demonstration sampler selects a random state s and chooses an action a from a Boltzmann distribution (softmax) over the learned q-values. The explanation sampler samples a random initial state s, a random initial action a, a random second action (different from the first) b, a random via state with non-zero reward (to fit the good or bad description) based on the discounted state visitation frequencies given a policy π. After ensuring that there is a next state to sample, the explanation sampler finally checks if the next state is good or bad comparing the q-values. The q-values provide information on whether, on the long run, a certain state or action will lead to better rewards. The

reward, demonstration or explanation samples generate natural language strings that take the forms of: “The reward in state 7 when performing action Left is -0.05.", “In state 7 you should perform action Up.", “In *state*, *action* is *better/worse* than *action* because it may eventually lead you through *state* and that is *better/worse*", respectively.

Procedure

We run a comparative study. Every 10 steps, for each of the three approaches, we: estimate the reward, compute the associated optimal policy π^* from the reward parameters w^*, and evaluate that policy in the correct MDP. We test the proposed approaches by providing the learner with 15 samples, using each of the three methods. Each sample is selected randomly according to the corresponding distributions. The results are depicted in Figure 6.8, and correspond to the performance of the policy of the learner using the learned reward in the original problem. Each plot corresponds to the average of 30 independent runs where, in each run, the parameters w^* are sampled randomly from a Gaussian distribution with mean 0 and unit standard deviation. For reproducibility, we set 40 random seeds.

Results

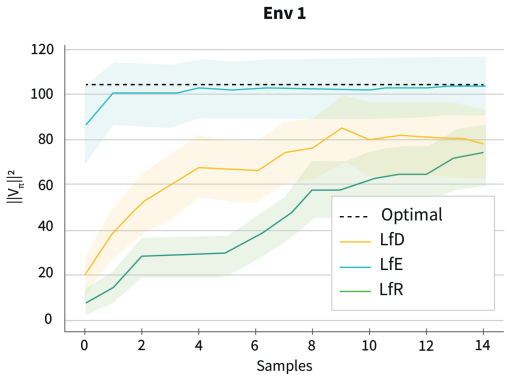

Figure 6.5: Results of Simulation with Env 1

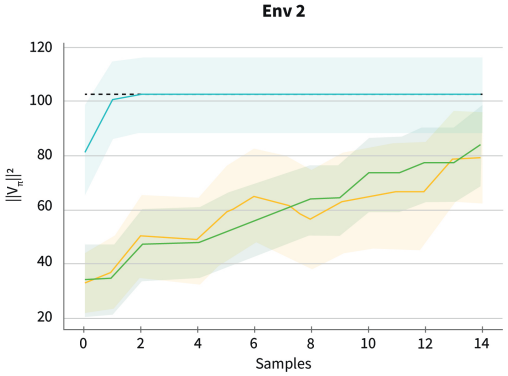

Figure 6.6: Results of Simulation with Env 2

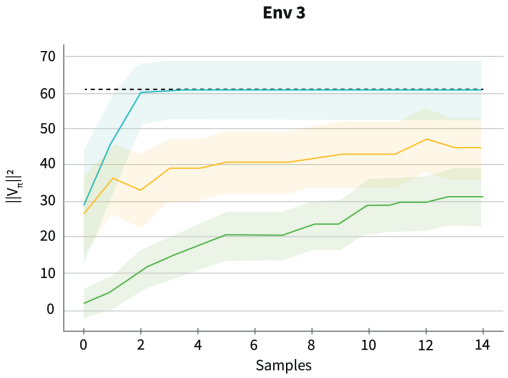

Figure 6.7: Results of Simulation with Env 3

Figure 6.8: Average return against the number of samples grouped by condition. Mean and confidence intervals for 40 seeds.

We plot the average return against the number of samples grouped by condition (Figure 6.8). The black dotted line corresponds to the expert's performance. As expected, and similar to results found by [156] when agents' were learning from descriptive feedback, the performance of the agent increases when learning from explanation samples. LfE agent outperforms both the LfD and LfR agents reaching near-optimal performance in all the environments. Looking closer to type of samples that hindered better performance, the analysis of the average returns after 1 learning update revealed a statistically significant difference between the performance of the LfE and LfD agents (Env 1: $t = -3.709$, $p = 0.000$, Env 2: $t = -3.495$, $p = 0.001$, Env 3: $t = -2.891$, $p = 0.007$). Moreover, the LfD agent performs better than the LfR agent in 6.1 and 6.2, and achieves similar performance in 6.2. All the agents eventually learn how to perform the task.

The results of the simulations validate **H1**, showing trends that the LfE approach leads to more efficient learning. That is, the LfE algorithm enables the learner to imitate the expert behavior achieving better performance than LfR and LfD do all over the tasks. This result confirms that LfE

is more sample efficient than LfR and LfD approaches in terms of the expert demonstration.

6.3 User Study

To validate whether this learning approach would be valuable in teaching scenarios involving humans, we evaluated the different teaching signals in a user study. The teaching signals generated with our system, take the form of sentences including the information detailed in section 6.2. We seek to investigate how humans select teaching signals and if their choice changes depending on the position of the learner with respect to the goal.

Motivated by findings from social sciences on the prominent role of explanations in human learning and inference [9], we hypothesize that when choosing among teaching signals **(H2)** humans will generally prefer explanations. Moreover, **(H3)** humans would prefer explanations over both rewards and demonstrations depending on the contextual situation, i.e., how close is the goal with respect to the learner position.

Methodology

As for the simulation experiment, for the user study we consider a gridworld environment consisting of 19 states and four objects 6.1. Participants are asked to play the role of the teacher and select the most appropriate teaching signal to help the learner navigate towards one of the four objects depending on the situation. The experiment consist of four phases: (1) familiarization, (2) structured teaching signals, (3) free-form explanations, and (4) subjective evaluation around the informativeness of the structure of the teaching signal. Participants had the possibility to navigate the environment until they were comfortable with it. Within the study, we propose four situations based on the goal (Figure 6.10), each situation including eight relevant states depicting the learner distance from the goal: far, ambiguous and adjacent. Far goals are equal or more than 3 steps away from the player, ambiguous goals are two steps away from the player and had a negative colored cell at equal distance, adjacent goals are next to the player. Situations were accompanied with information about the goal as well as a set of structured teaching signals. Examples of structured teaching signals include: (1) demonstration, i.e.,*"In cell 17 you should move right."*, (2) reward, i.e.,*"The amount of points you get in cell 17 when moving left is around 0.05."*, (3) explanation, i.e.,*"In cell 17, moving Right is better than Up because it may eventually lead you through cell 11 and that is good."* Finally, we ask three exploratory questions to rate the informativeness of the general structure of the explanations we employ throughout the study.

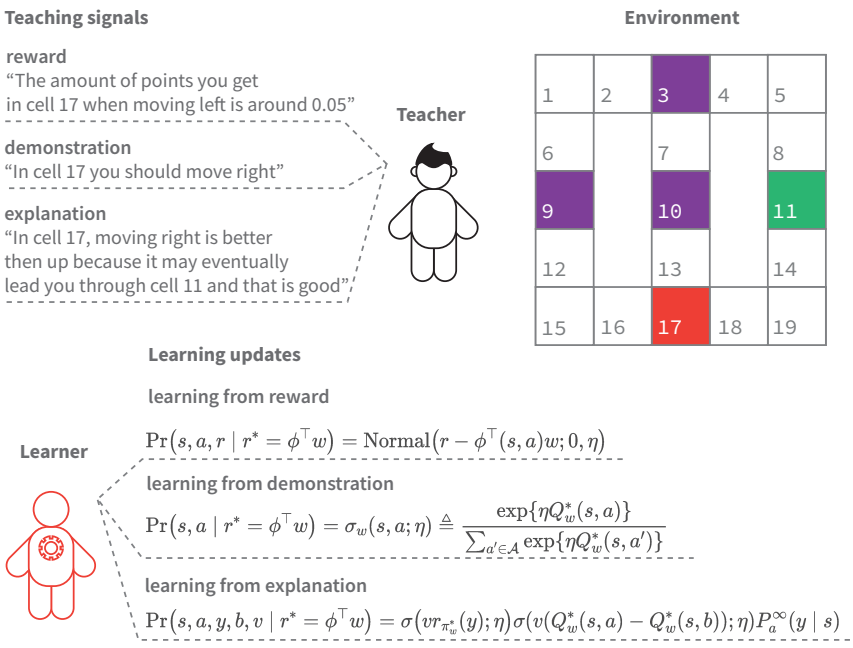

Figure 6.9: An example of three teaching signals for the depicted situation.

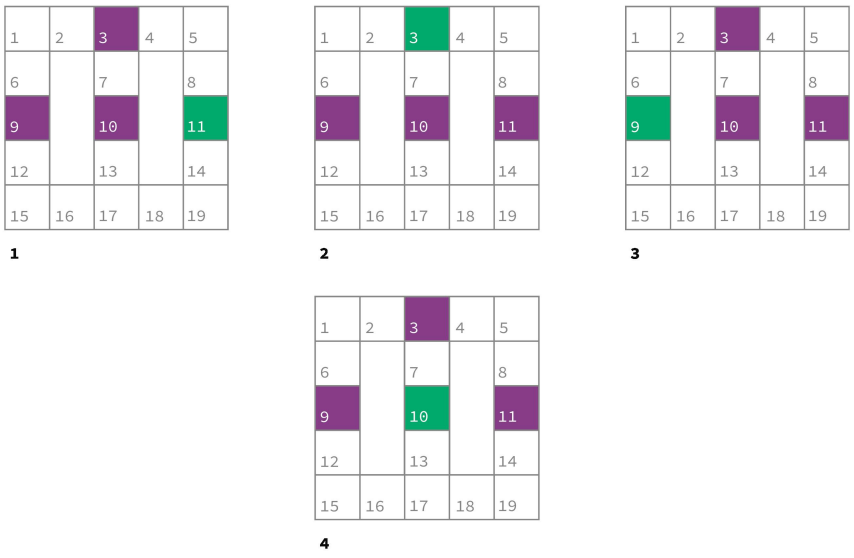

Figure 6.10: Navigational environment used for the user study

To measure humans’ explanatory preferences **(H2)**, we asked participants to select the best teaching signal among demonstration, reward, and explanation, and to provide an open answer to the question *“If you were asked to provide an explanation in this situation, what would that be?*". To measure whether the contextual situation affects human’s explanatory preferences **(H3)** we

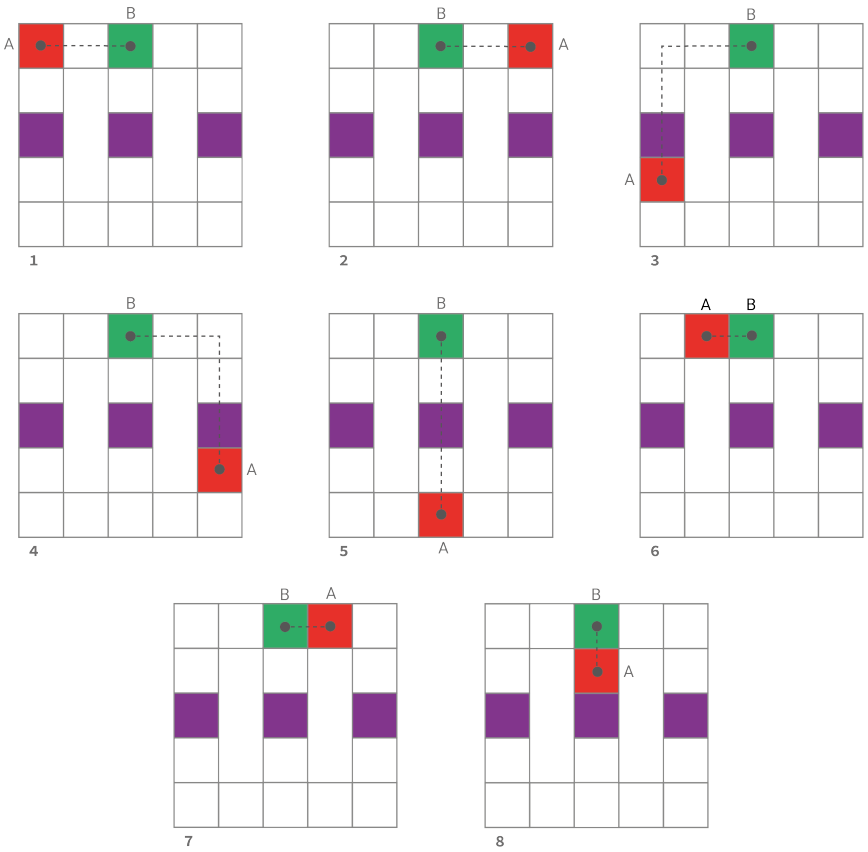

Figure 6.11: An example of eight positions of the learner (red checker) with respect to the goal

cluster the situations based on the learner's distance from the goal.

Participants

We recruited 150 participants using Prolific[1]. All participants were English speakers and gave informed consent to participate (age M = 40.24, SD = 13.43, gender Female = 91, Male = 59). We introduced some attention and verification questions in order to ensure the quality of the data. We asked four multi-answer questions related to the scenario (e.g., *"What is the goal of the player? (Check all that apply)*") and computed an attention score based on the number of correct answers. The criteria to exclude participants were: not having completed the entire experiment; and having an attention score of less than 70%. Consequently, we ran the data analysis on both the entire sample and on the reduced sample of 60 more reliable participants (age M = 39.83, SD = 12.85, Female = 33).

[1] `https://prolific.co/`

Materials

The self-assessed questionnaire included some demographic questions (age, gender, higher level of education), two items regarding participants self-perceived familiarity with navigational games and with reinforcement learning, three items regarding the perceived informativeness of the proposed teaching signals and four validation questions randomly dispersed in the questionnaire to evaluate their understanding of the rules of the game.

Procedure

After replying to the self-assessed questionnaire, participants were asked to consider a single-player video game based on a two-dimensional navigational environment in which a player, represented by a red checker, has to reach a hidden goal. Their role was to guide an hypothetical learner by providing the most informative teaching signals by either choosing among the structured signals and/or writing their own feedback.

Results

General Preference To obtain an overview of the participants' explanatory preferences, we summed up the number of teaching signals in all situations. The one-way analysis of variance of the teaching signals shows that there is a significant difference on how participants selected demonstrations, rewards and explanations (Kruskal-Wallis H test: $H = 16.810, p = .000$). The Mann-Whitney U test between explanations ($M = 13.375, SE = 5.909, SD = 24.763$) and demonstrations ($M = 4.82, SE = 6.120, SD = 33.858$) revealed a significant difference ($U = 10.0; p = .02$). A significant different was also found between demonstrations and rewards ($M = 1.625, SE = .374, SD = 5.402$)($U = 64, p = .00$), and explanations and rewards ($U = 61, p = .002$). These results suggest that participants' generally prefer demonstrations.

Influence of Contextual Situation The Kruskal-Wallis H test of the teaching signals, i.e., explanations, demonstrations, rewards, revealed that the main effect of the situation was significant across situation Situation 3 ($H(1, 150) = 13.022, p = .001$)), and Situation 4 ($H(1, 150) = 16.810, p = .000$). The specific values per each teaching signal were: Situation 3 [explanations ($M = 12.75, SE = 3.648, SD = 9.653$), demonstrations ($M = 37.875, SE = 7.024, SD = 18.583$), rewards ($M = 1.875, SE = 1.315, SD = 3.479$)] Situation 4 [explanations ($M = 13.375, SE = 5.909, SD = 9.653$), demonstrations ($M = 45.0, SE = 6.120, SD = 18.583$), rewards ($M = 1.625, SE = .374, SD = 3.479$)].

The analysis of variance in teaching signals revealed a main effect of the player distance from the goal for adjacent goals in Situation 2 ($H(3, 150) = 4.705, p = .049$) and in all situations for far goals: Situation 1 ($H(3, 150) = 4.705, p = .006$), Situation 2 ($H(3, 150) = 4.705, p = .025$), Situation 3 ($H(3, 150) = 2.0, p = .014$), Situation 4 ($H(3, 150) = 4.705, p = .02$). These results indicate that the teaching signals are selected differently based on the position of the player with respect to the goal.

Overall, participants reduce the number of explanations when the player is adjacent to the goal, while consistently choose to give more explanations when the player is far from the goal.

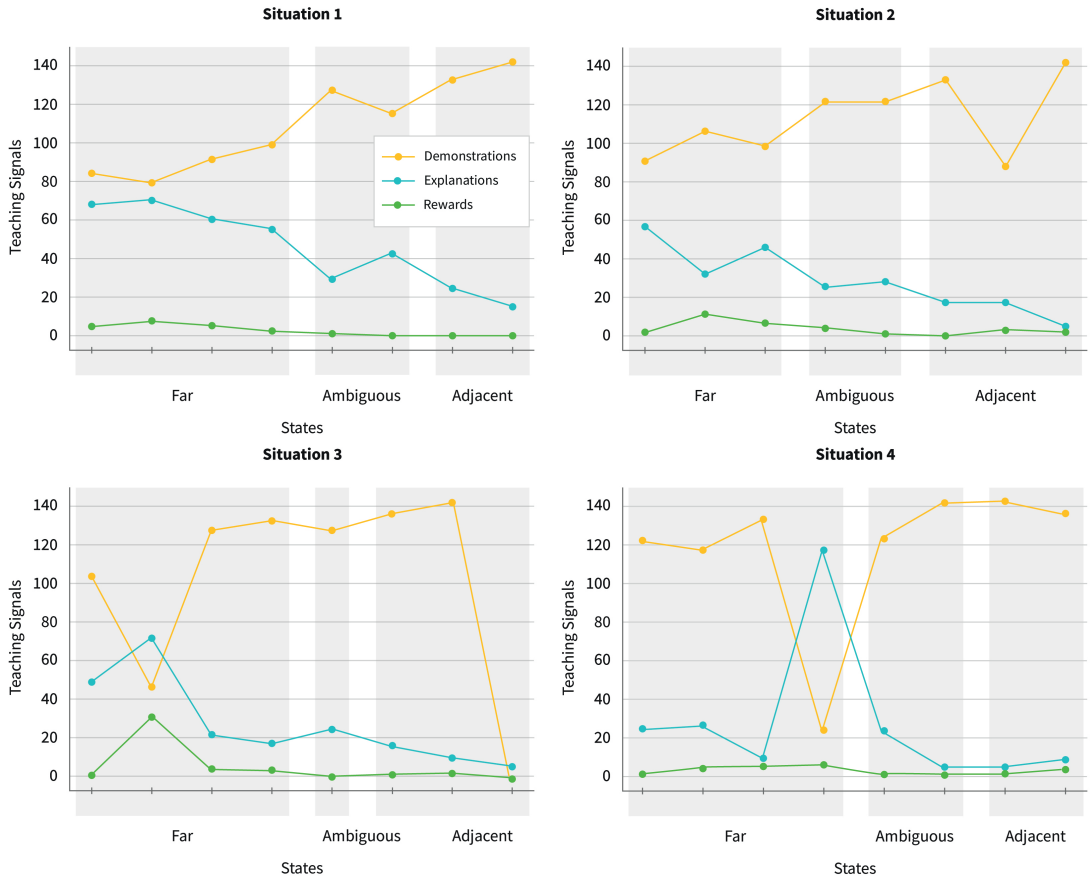

Figure 6.12: Number of teaching signals per situations against position of the learner.

Our results show that participants prefer to teach through demonstrations significantly more than rewards and explanations, therefore **H2** is not confirmed. Moreover, in all situations participants consistently tend to decrease the number of explanations as the learner get closer to the goal, confirming **H3**.

6.4 Findings

In this work we explored the role of explanation in learning and teaching tasks. First, we investigated whether or not explanations can lead to better performance and introduced the *learning from explanations* (LfE) approach for recovering a reward function from explanations of another agent. Second, we evaluated the generated explanations in a user study. Hence, we compared the performance of an agent learning from explanations against demonstrations and rewards.

6.4 Learning Performance

Regarding **H1** we predicted that agents would learn more efficiently from explanations than other types of teaching signals. The simulations confirmed our hypothesis by showing that in the explanation condition the agent was capable of achieving better performance. With our approach rewards were chosen to maximize the likelihood of the data given as a set of traces of optimal behavior, allowing us to combine a supervised-learning component with a flexible hypothesis class given as input. The LfE algorithm is simple, with relatively low computational cost per iteration. However, the maximum likelihood algorithm generally behaves well when the expert's demonstration are representative of the task [209]. Further work should focus on testing this approach in other learning scenarios adopting more robust and scalable implementations.

6.4 Teaching Signals

According to **H2**, we expected humans to generally prefer explanations over rewards and explanations to guide a player towards a hidden goal. However, this was not supported. Instead, we found that participants preferred to teach through demonstrations significantly more than rewards and explanations. Our results are in line with the work of [39] which compares *simplicity* and *probability* in causal explanation, and states that simpler explanations are preferred and judged more likely. Thus, teaching signals which invoke a limited number of causes whilst still conveying relevant aspects of a task are favored, i.e., demonstrations.

6.4 Contextual Situation

In **H3**, we hypothesized that humans would choose explanations depending on the contextual situation. We observed that in all situations participants consistently tend to decrease the number of explanations as the player get closer to the goal. This might be due the fact that explanations

conveys complex concepts better than demonstrations while relying on shared context to permit high bandwidth. In contrast, demonstrations are lower-bandwidth but more robust [214], therefore seen as more useful in situations in which there is little ambiguity, i.e., when the learner is not far from the goal.

6.5 Conclusion

Throughout this work we explore the problem of teaching and learning from explanations and provide a framework to compare learning from explanations with learning from other types of teaching signals. We present an application of maximum likelihood inverse reinforcement learning to the problem of training an agent to follow different teaching signals representing high-level tasks. We evaluate our approach in three navigational scenarios. We then undertake a user study with 150 participants to investigate humans' preferences between the different types of teaching signals and the impact of contextual situations on their choice. The first takeaway from this work is that we can improve agents performance by integrating explanations into IRL. The second takeaway, derived from the user study, is that in the context of interactive task learning, humans might prefer different types of teaching signals depending on the contextual situations. To evaluate if our approach would work in human-AI partnership, future studies should evaluate the machine-generated explanations on human learning. Moreover, given the interactive nature of the explanatory process and the large variability of human explanations, a more advanced systems that is able to incorporate human free-form explanations, instead of structured explanations, should be implemented and tested in the wild.

As a consequence of our approach we show that explanations can be a more succinct, robust, and transferable way to represent tasks. This approach is presumably robust enough to be applicable to a large range of sequential planning tasks in both human-agent and agent-agent settings.

Chapter 7

Future Work and Conclusion

There is a growing interest in the research artificial intelligence community to augment human-AI partnership. Research in this direction stresses that either fully autonomous systems and manual approaches, i.e., approaches that require the human to create the systems' rules manually, have limitations that human-AI partnership could overcome. Despite the strong predictive performance of AI systems and the considerable advances toward more robust and adaptive manipulation, perception and planning in robots, fully autonomous systems are often not desirable due to safety, ethical, and legal concerns [215, 216]. On the contrary, manual approaches can be inaccurate and time consuming [217, 218]. Existing efforts to enable human-AI partnership highlight the importance of developing AI systems that are able to explain their inner workings and learn through natural interaction with humans [5]. While AI's explanations could provide the humans with additional clues about the functioning and solutions of AI systems, human's explanations could help the AI systems to reason upon relevant alternatives, and thus optimize their learning process.

Existing explainability methods (1) neglect the effects of AI system's explainability on human cooperative behaviors (2) are not designed to provide explanations about suboptimal actions for the human to learn from the algorithm, and (3) are not built for more efficiently transferring knowledge among agents. We bridge this gap by evaluating explainability in human-agent teams and human decision-making, and by building computational models to enable AI systems to both provide explanations about the suboptimality of their actions and to learn from contrastive explanations of an expert.

7.1 Effects of Agents' Explanations on Teamwork

- **RQ1** Does explaining the strategies of agents in human-agent teams foster more collaborative behaviors in the human?

Previous research on explanations in human-AI partnership was primarily concerned with enabling AI systems to provide a human partner information to develop an accurate mental model of the system and its behavior [184, 219]. In this context, particular attention has been placed on trust calibration, and task performance while other aspects of teamwork, such as cooperative behaviors, have been ignored [220]. Motivated by the fact that transparency about choices tends to lead to an increase in contributions and collusion [160, 221], we decided to explore the effects of artificial agent's explanations on human cooperative choices. Therefore we implemented a transparency module in a collaborative game scenario with a mixed human-agent team. We observed how the strategy and explainability of artificial agents influence human cooperative

behavior in teamwork. Within the limits of the results found, we observed significant effects of explainability on trust, group identification and human likeness. In particular, we showed that adding transparent behaviour to an unconditional cooperator negatively affects the perceptions people have of the artificial partners. This aspect has interesting implications in the context of public goods games and the design of relational and social capabilities in intelligent systems.

Further research should take into consideration the preexisting social value orientations in participants (i.e., prosocial, individualistic, and competitive orientations) [222] and randomize the sample among the experimental conditions before running the study.

During collaborations, humans communicate intent with different methods. This communication can be verbal and non-verbal, potentially spanning multiple levels of abstractions [216, 23]. Other types of explainability methods, including non-verbal cues and multimodal explanations, should be explored.

Although the transparency module we developed in this thesis allowed the agents to explain their strategies during the game-play, this module did not account for the model of the human collaborator. A logical extension of the work presented in this thesis would be to implement a transparency module that adapts the explanations of the agents' strategy according to the cooperative behavior of the human, optimizing for the number of human cooperative choices.

Another interesting research direction would be to explore the agents' strategy explanations in adversarial settings where the human and the AI systems have contrasting goals. In this context, explainability would be a possible way to mitigate these conflicts.

7.2 Explainable Agency by Revealing Suboptimality

- **RQ2** How can intelligent autonomous agents provide explanations about their behaviors to enhance human understanding of a new task?

The primary function of explanations is to facilitate learning [9, 223]. Explanations help to establish a connection between some event that has occurred, and the causes of said event. This allows for generalisation of these causes and effects to other contexts. Consequently, explanations scaffold causal learning and have a crucial role in inference. As people find contrastive explanations, i.e., explanations that explain the cause of an event relative to some other event, more intuitive and more valuable than other kinds of explanations [29], implementing contrastive explanations is a good way to approach the problem of designing explainable intelligent agents.

Existing work on designing explainable intelligent agents focuses on comparing possible plans, and including justification as to why one plan is chosen over an alternative. However, this work

often does not account for whether the course of action chosen by the agent was, in fact, the optimal action for the given scenario. Furthermore, few examples in the literature of autonomous and explainable systems are tested in a child-robot interaction scenario [224, 225].

Studying explanations in the context of children's learning raises a number of interesting methodological and technical challenges, such as (1) how to design the interaction between the child and the robot to achieve better learning outcomes, e.g., roles, learning objectives (2) how to adapt the difficulty of the learning task to ensure the right level of engagement of the child, and (3) how to build a system that is autonomous yet robust enough for interacting with children.

Although scholars have explored psychological processes which surround the design of child-robot educational tasks [226, 42, 227, 228], and have shown that robot's help-seeking intervention strategies may shape children's suboptimal help-seeking behaviors and subsequently influence their learning outcomes [229], little work investigates the effects explaining the sub-optimal actions of a robot on children learning.

We strove to explore the topic of explaining the suboptimality of robot's actions to make children better understand a new task and improve their logical and mathematical thinking. We adapted a search-based approach for generating contrastive explanations and evaluated the validity of our approach in an experiment involving seven-year-old children. We could not find an effect of explainable agency on neither children's efficiency of playing the game, nor on the children's perception of the robot. However, we showed that the children in the explainable condition reported a significantly lower perceived difficulty in performing the post-test in respect to the pre-test. These results underline that providing explanations about suboptimal behaviors positively affects learning scenarios with children.

Recent approaches to plan explanation stress the importance of moving beyond the explanation as a soliloquy and framed the explanation, or model update, as a model reconciliation problem [85, 97]. Future work on human learning from machine-generated explanations should consider the adaptation of the explanation to the learner's mental model of the task [230]. An example of the child-robot scenario we have developed would be taking into account the level of expertise of the children for generating appropriate explanations. The robot should provide less or different types of information to the children that play better (e.g. less number of moves for solving the game). For those that encounter problems, the robot would explain details and alternatives about possible solutions using hierarchical terms at different levels of abstraction.

Another challenging future direction is using questions to explain the robot's knowledge implicitly. This would be achieved by building an interactive task learning scenario in which the robot learns from the child. In the context of the proposed game scenario, the robot could

query the child about the rules, the goal, or the optimal policy to play the game. The explainable robot could improve the learning experience by revealing to the child, teacher, or peer what is known and what is unclear [93]. By phrasing questions in specific ways, the robot could provide information about the learning task, and foster the children to learn while trying to demonstrate or explain possible solutions to the planning problem.

7.3 Learning from Explanations as Inverse Planning

- **RQ3** How can intelligent autonomous agents learn from another agent's explanations?
- **RQ4** Does explanation compared to demonstration and reward signals lead to better learning?

Research on incorporating human expertise and knowledge in machine learning proposes methods to allow the systems to learn from observing the human behavior [15], extracting relevant information from dialogue [16, 17], explicitly receiving instructions, demonstrations [18, 19], and feedback [20, 21]. Recent work underlines the importance of endowing AI systems with the ability to learn using counterfactuals, i.e., contrast cases, [68], and causal models [231].

We argued that explanations are a valuable way to transfer knowledge among agents and to incorporate explanations into maximum likelihood inverse reinforcement learning. The proposed framework enabled us to evaluate learning from explanations against learning from other teaching signals coming from an expert agent.

In our work, we designed contrastive explanations that compare the valence of actions and states, i.e., how good/bad they are with respect to other actions and environmental states. Yet explanations can take several other forms. Explanations can vary between human teachers and with respect to their assumptions about the learner's knowledge and learning capabilities [232]. For example, the type of information that an explanation holds varies with the human teachers' familiarity with the task or their model of the agent's functioning. Explanations can also result from a dialogue [136, 11]. In this context, the complexity and heterogeneity of the interacting parts requires multiple iterations to enable both humans and AI systems to model their counterparts. Yet, an explanation is oftentimes the result of segmented information coming from both the *explainer* and the *explainee*. Such diversity in explanations are omnipresent in the real world and it would be interesting to incorporate them in future studies.

While the majority of the Explainable AI literature envisages a one-shot type of interaction, attempts to tackle the problem of interactive explanation over long-term interactions are few. Future work should focus on understanding what the facets of explainability in interactive

scenarios are. Moreover, most of the existing explainability methods, both in interpretable machine learning and explainable agency, are associational [45]; they depend on only data observations and do not employ any form of causal reasoning [233]. Thus, while they can provide some insight on the models used by AI-enabled systems, e.g., what key features of the observation(s) led to the system's prediction, they cannot answer some types of questions that users may have. For example, particularly counterfactual questions, e.g., what would be observed if the environment is changed in a specific manner, what caused the observation(s) to occur. Future research should seek to understand the state-of-the-art in causal methods for explainable agency, their limitations, and promising approaches for solving them. We believe that both psychological analyses of how humans formulate explanations and philosophical analyses of the fundamental nature of causation [234, 235] may serve as foundations for developing causal methods for building explainable agency.

Future works should study different methods to learn from explanation, for example, by testing different combinations of distributions and statistical inference (i.e., Bayesian inference) [236] and counterfactual reasoning. Lastly, we advocate the importance of studying the effect of machine-generated explanations on human learners. An interesting future direction would be to test how our machine-generated explanations would affect human performance in a learning task and how human free-form explanations could be used to train an agent. Learning often happens through multiple interactions and between multiple learners. Another valuable extension of the proposed computational models is the integration of explanations in interactive scenarios involving multiple agents and teams.

7.4 Conclusion

Developing a two-way dialogue between humans and decision-making systems is crucial for leveraging the relative strengths of humans and AI systems. Such a dialogue requires an understanding of the perceptual and representational spaces that characterize both humans and AI systems [5].

Alongside the refinement of explainability desiderata and methods, recent work on explainability has shown a renewed interest toward social sciences [29]. Explainability has reached a growing number of niche areas of research (e.g., embodied agents [12], reinforcement learning [11]). We are slowly moving towards standardized definitions and evaluation metrics [237]. The work done so far has explored different facets of explanations in human-AI collaborative scenarios by addressing four main research goals.

- The primary research goal was to provide a definition of explainability which acknowledged the challenges related to both building explainability for sequential decision making agents and embodied agents. We showed that explainability in sequential decision making agents inherently differs from explainability in prediction models due to the dependency among the variables affecting the decision-making process, e.g., states and actions. Furthermore, we argued that when intelligent agents possess an embodiment, the large range of communication modalities they can access requires a change in how explainability is designed and evaluated. As a result, we reviewed existing literature on the topic of explainable embodied agents and analyzed definitions, implementations, and evaluation metrics.

- The second research goal was to investigate how explanations affect human-agent teams, in particular analysing the effects of revealing the artificial agents' strategies on human cooperative behaviors. This research goal raised interesting challenges related to how explainability is exploited in mixed human-agent and multi-agent settings and what other aspects, besides trust calibration and team efficiency, should be considered. We implemented a transparency module in mixed human-agent teams and highlighted that, in collaborative scenarios, explainability does not directly link with human cooperative behaviors and that its effects change with respect to the strategy adopted by the artificial agents.

- The third research goal was to develop computational mechanisms that can enable AI systems to increase the effectiveness of the human understanding of a new task. This research goal involved (1) identifying an interaction scenario in which the explanations of the decision-making process of an AI system might be relevant for the human, (2)

implementing an autonomous system that can provide information about aspects of its decision-making that are important from a human learning point of view, (3) deploying and evaluating the system with humans. We designed a system that uses a minmax algorithm to explain the suboptimality of a robot actions and deployed such system in a learning scenario involving children. We argued that the AI system providing explanations about its suboptimal actions allows the human to reason about the contrastive cases, thus learning the best way to proceed without being overwhelmed.

- The fourth research goal was pivoted on computational mechanisms that an AI system apply to learn more efficiently from an expert. This research goal brought up challenges related to (1) how to incorporate the expert's knowledge in a more concise representation of a task, i.e., contrastive explanations, (2) how to make a machine learning model reason upon such representation, and (3) how to evaluate such approaches against other existing direct exploration and imitation learning approaches. We addressed these challenges by integrating contrastive explanations in maximum likelihood inverse reinforcement learning and by providing a framework to compare learning from different types of teaching signals.

Final Remarks

The ability to cooperate and communicate with others is the key to our success and survival [238]. Finding strategies to understand the reasoning of our teammate(s) and act accordingly implies being able to learn about and from them. Thus, learning holds a prominent role in adapting to new situations. To further narrow the problem, expressing our experience in a precise and succinct way enhances our ability to understand a new problem and learn how to deal with it. On one side, by conveying meaningful information about our reasoning process through explanations, we increase our chances to retain what we have learned and allow others to benefit from it. On the other side, by being able to make sense of others' explanations we tap into their experience and move forward faster. Thanks to others' explanations, we explore alternatives to what we experience firsthand and unleash our capacity to grow as human beings rather than being confined to genetically coded and relatively fixed abilities.

Language and other forms of natural communication enable us to share the results of our knowledge with others. We culturally evolved by learning from others. We are adaptive learners who, even as infants, carefully select when, what, and from whom to learn [238].

We use causes and effects to organize our knowledge of the world. We observe and detect regularities, i.e., association, we act and choose among deliberate alternatives, choosing the one most likely to lead to the desired outcome, i.e., intervention. Finally, we understand something

and retrospect about possible alternatives, i.e., imagination [62].

Through building computational models for explanation-guided learning specifically for human-AI partnership, we advanced our understanding of explainability in collaborative scenarios in mixed human-agent teams. Furthermore, we endowed intelligent agents with the ability to explain their inner workings in a way that is relevant both for making them transparent to humans and augmenting human learning. Finally, we extended the learning abilities of intelligent agents by enabling them to retrospect on possible alternatives and learn from human compatible teaching signals, i.e., contrastive explanations. We argue that this is a first step towards building the explainable agency of artificial intelligent partners, and a first step towards improving the learning process of both humans and AI systems through reflection upon causes and effects.

Bibliography

[1] B. Malle. How the mind explains behavior: Folk explanations, meaning, and social interaction. In -, 2004.

[2] J.-M. Fellous, G. Sapiro, A. Rossi, H. S. Mayberg, and M. Ferrante. Explainable artificial intelligence for neuroscience: Behavioral neurostimulation. *Frontiers in Neuroscience*, 13, 2019.

[3] C. Rudin. Algorithms for interpretable machine learning. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

[4] O. X. Kuiper, M. van den Berg, J. van den Burgt, and S. Leijnen. Exploring explainable ai in the financial sector: Perspectives of banks and supervisory authorities. *ArXiv*, abs/2111.02244, 2021.

[5] B. Kim. Interactive and interpretable machine learning models for human machine collaboration. In *PhD Thesis*, 2015.

[6] S. J. Russell. Human-compatible artificial intelligence. *Human-Like Machine Intelligence*, 2021.

[7] D. Gunning. Darpa's explainable artificial intelligence (xai) program. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019.

[8] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. ISSN 15662535. doi: 10.1016/j.inffus.2019.12.012. URL `http://arxiv.org/abs/1910.10045`.

[9] T. Lombrozo. The structure and function of explanations. *Trends in Cognitive Sciences*, 10 (10):464–470, 2006. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2006.08.004. URL `https://www.sciencedirect.com/science/article/pii/S1364661306002117`.

[10] G. Vilone and L. Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion*, 76:89–106, 2021.

[11] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere. Explainable reinforcement learning through a causal lens. *ArXiv*, abs/1905.10958, 2020.

[12] S. Wallkotter, S. Tulli, G. Castellano, A. Paiva, and M. Chetouani. Explainable agents through social cues: A review, 2020.

[13] P. Lertvittayakumjorn and F. Toni. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528, 2021.

[14] J. E. Laird, K. A. Gluck, J. R. Anderson, K. D. Forbus, O. C. Jenkins, C. Lebiere, D. D. Salvucci, M. Scheutz, A. L. Thomaz, J. G. Trafton, R. E. Wray, S. Mohan, and J. R. Kirk. Interactive task learning. *IEEE Intelligent Systems*, 32:6–21, 2017.

[15] Y. Lashkari, M. Metral, and P. Maes. Collaborative interface agents. In *AAAI*, 1994.

[16] B. Kim, C. M. Chacha, and J. A. Shah. Inferring robot task plans from human team meetings: A generative modeling approach with logic-based prior. *ArXiv*, abs/1306.0963, 2013.

[17] Z. Zhang, R. Takanobu, M. Huang, and X. Zhu. Recent advances and challenges in task-oriented dialog system. *ArXiv*, abs/2003.07490, 2020.

[18] B. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics Auton. Syst.*, 57:469–483, 2009.

[19] K. Judah, S. Roy, A. Fern, and T. G. Dietterich. Reinforcement learning via practice and critique advice. In *AAAI*, 2010.

[20] A. L. Thomaz and C. Breazeal. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *AAAI*, 2006.

[21] W. B. Knox and P. Stone. Interactively Shaping Agents via Human Reinforcement: The TAMER Framework. In *Proceedings of the Fifth International Conference on Knowledge Capture*, K-CAP '09, pages 9–16, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-658-8. doi: 10.1145/1597735.1597738. URL `http://doi.acm.org/10.1145/1597735.1597738`.

[22] S. V. Albrecht and P. Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *ArXiv*, abs/1709.08071, 2018.

[23] A. D. Dragan, K. C. T. Lee, and S. Srinivasa. Legibility and predictability of robot motion. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308, 2013.

[24] H. M. Wellman. Reinvigorating explanations for the study of early cognitive development. *Child Development Perspectives*, 5:33–38, 2011.

[25] P. Thagard. *Explanatory coherence*. Cambridge University Press: Behavioral and Brain Sciences, 1993.

[26] D. S. Krull and C. A. Anderson. *Explanation, Cognitive Psychology of*. Current Directions in Psychological Science, 2001.

[27] D. Lewis. Causal explanation. In D. Lewis, editor, *Philosophical Papers Vol. Ii*, pages 214–240. Oxford University Press, 1986.

[28] M. M. A. de Graaf and B. Malle. How people explain action (and autonomous intelligent systems should too). In *AAAI Fall Symposia*, 2017.

[29] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, abs/1706.07269, 2017. URL `http://arxiv.org/abs/1706.07269`.

[30] S. Chin-Parker and A. Bradner. Background shifts affect explanatory style: how a pragmatic theory of explanation accounts for background effects in the generation of explanations. *Cognitive Processing*, 11:227–249, 2009.

[31] A. Aliseda. *Abductive reasoning*, volume 330. Springer, 2006.

[32] D. G. Campos. On the distinction between peirce's abduction and lipton's inference to the best explanation. *Synthese*, 180(3):419–442, 2011. ISSN 00397857, 15730964. URL `http://www.jstor.org/stable/41477565`.

[33] C. G. Hempel and P. Oppenheim. Studies in the logic of explanation. *Philosophy of Science*, 15:135 – 175, 1948.

[34] T. Lombrozo. Explanation and abductive inference. In *The Oxford Handbook of Thinking and Reasoning*, 2012.

[35] J. F. Woodward and W. Salmon. Scientific explanation and the causal structure of the world. *Noûs*, 22:322, 1984.

[36] T. Lombrozo and S. Carey. Functional explanation and the function of explanation. *Cognition*, 99(2):167–204, 2006. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2004.12.009. URL https://www.sciencedirect.com/science/article/pii/S0010027705000466.

[37] I. S. Newton. *Philosophiæ naturalis principia mathematica*. Imprimatur S. Pepys, 1973. URL http://cudl.lib.cam.ac.uk/view/PR-ADV-B-00039-00001/24.

[38] F. J. Anscombe and R. J. Aumann. A definition of subjective probability. *The Annals of Mathematical Statistics*, 34(1):199–205, 1963. ISSN 00034851. URL http://www.jstor.org/stable/2991295.

[39] T. Lombrozo. Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3): 232–257, 2007. ISSN 0010-0285. doi: https://doi.org/10.1016/j.cogpsych.2006.09.006. URL https://www.sciencedirect.com/science/article/pii/S0010028506000739.

[40] L. Davachi, J. P. Mitchell, and A. D. Wagner. Multiple routes to memory: Distinct medial temporal lobe processes build item and source memories. *Proceedings of the National Academy of Sciences of the United States of America*, 100:2157 – 2162, 2003.

[41] C. C. Chase, D. B. Chin, M. A. Oppezzo, and D. L. Schwartz. Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of Science Education and Technology*, 18:334–352, 2009.

[42] S. Chandra, P. Dillenbourg, and A. Paiva. Children teach handwriting to a social robot with different learning competencies. *International Journal of Social Robotics*, 12:721–748, 2020.

[43] Z. C. Lipton. The mythos of model interpretability. *Queue*, 16:31 – 57, 2018.

[44] Y. Lien and P. W. Cheng. Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, 40:87–137, 2000.

[45] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[46] A. Heuillet, F. Couthouis, and N. D. Rodríguez. Collective explainable ai: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values. *IEEE Computational Intelligence Magazine*, 17:59–71, 2022.

[47] T. Huber, D. Schiller, and E. André. Enhancing explainability of deep reinforcement learning through selective layer-wise relevance propagation. In *KI*, 2019.

[48] P. Madumal. Explainable agency in intelligent agents: Doctoral consortium. In *AAMAS*, 2019.

[49] P. Langley. Explainable agency in human-robot interaction. In *Association for the Advancement of Artificial Intelligence*, 2016.

[50] P. Langley, B. Meadows, M. Sridharan, and D. Choi. Explainable agency for intelligent autonomous systems. In *Twenty-Ninth IAAI Conference*, 2017.

[51] Y. Zhang, H. Zhuo, and S. Kambhampati. Plan explainability and predictability for cobots. *ArXiv*, abs/1511.08158, 2015.

[52] P. Langley. Explainable, normative, and justified agency. In *AAAI*, 2019.

[53] S. Kambhampati. Challenges of human-aware ai systems. *ArXiv*, abs/1910.07089, 2019.

[54] J. Hoffmann and D. Magazzeni. Explainable ai planning (xaip): Overview and the case of contrastive explanation (extended abstract). In *Reasoning Web*, 2019.

[55] S. Rathi. Generating counterfactual and contrastive explanations using shap. *ArXiv*, abs/1906.09293, 2019.

[56] D. Amir and O. Amir. Highlights: Summarizing agent behavior to people. *AAMAS*, /, 2018.

[57] I. Lage, D. Lifschitz, F. Doshi-Velez, and O. Amir. Exploring computational user models for agent policy summarization. *IJCAI : proceedings of the conference*, 28:1401–1407, 2019.

[58] B. Hayes and J. Shah. Improving robot controller transparency through autonomous policy explanation. *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 303–312, 2017.

[59] A. Lindsay. Towards exploiting generic problem structures in explanations for automated planning. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 235–238, 2019.

[60] S. Krening, B. Harrison, K. M. Feigh, C. L. Isbell, M. Riedl, and A. Thomaz. Learning from explanations using sentiment and advice in rl. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1):44–55, 2017.

[61] B. Harrison, U. Ehsan, and M. O. Riedl. Guiding reinforcement learning exploration using natural language. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi Agent Systems*, AAMAS '18, page 1956–1958, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[62] J. Pearl and D. Mackenzie. The new science of cause and effect. In *The Book of Why*, 2018.

[63] A. McEleney and R. M. J. Byrne. Counterfactual thinking and causal reasoning. In *Thinking and Reasoning*. Taylor & Francis (Routledge), 2000.

[64] T. Miller. Contrastive explanation: A structural-model approach. *ArXiv*, abs/1811.03163, 2018.

[65] A. Kean. A characterization of contrastive explanations computation. In *PRICAI*, 1998.

[66] I. Stepin, J. M. Alonso, A. Catalá, and M. Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[67] R. M. J. Byrne. Counterfactual thought. *Annual review of psychology*, 67:135–57, 2016.

[68] R. Silva. Counterfactual MDPs: Planning Beyond Direct Control. In *PhD Thesis*, 4 2021. doi: 10.1184/R1/14428145.v1. URL `https://kilthub.cmu.edu/articles/thesis/Counterfactual_MDPs_Planning_Beyond_Direct_Control/14428145`.

[69] R. Borgo, M. Cashmore, and D. Magazzeni. Towards providing explanations for ai planner decisions, 2018.

[70] M. Cashmore, A. Collins, B. Krarup, S. Krivic, D. Magazzeni, and D. Smith. Towards explainable ai planning as a service. *arXiv preprint arXiv:1908.05059*, 2019.

[71] S. Tsirtsis, A. De, and M. G. Rodriguez. Counterfactual explanations in sequential decision making under uncertainty. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=M5h1l1SldlF`.

[72] S. Verma, J. P. Dickerson, and K. Hines. Counterfactual explanations for machine learning: A review. *CoRR*, abs/2010.10596, 2020. URL `https://arxiv.org/abs/2010.10596`.

[73] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021. doi: 10.1109/ACCESS.2021.3051315.

[74] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, and F. Doshi-Velez. Explainable reinforcement learning via reward decomposition. In -, 2019.

[75] R. Cardoso. *Generation of Explanations in Reinforcement Learning*. Fenix Técnico Ulisboa, 2019.

[76] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan. Enabling Robots to Communicate Their Objectives. *Autonomous Robots*, 43(2):309–326, jul 2019. ISSN 15737527. doi: 10.1007/s10514-018-9771-0. URL http://dx.doi.org/10.15607/RSS.2017.XIII.059http://www.roboticsproceedings.org/rss13/p59.pdf.

[77] Y. Amitai and O. Amir. "i don't think so": Disagreement-based policy summaries for comparing agents. *ArXiv*, abs/2102.03064, 2021.

[78] A. C. Scott, W. J. Clancey, R. Davis, and E. H. Shortliffe. Explanation capabilities of production based consultation systems. In *ACL Microfiche Series 1-83, Including Computational Linguistics*, 1979.

[79] D. Warren. Implementing Prolog-Compiling Predicate Logic Programs. Technical report, Technical Report 39 and 40, Dept. of Artificial Intelligence, Univ. of Edinburgh., Edinburgh, 1977.

[80] W. van Melle. Mycin: a knowledge-based consultation program for infectious disease diagnosis. *International Journal of Human-computer Studies International Journal of Man-machine Studies*, 10:313–322, 1978.

[81] M. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages*, volume 1555, pages 1–10, Paris, 1999. Springer. ISBN 3540657134. doi: 10.1007/3-540-49057-4_1.

[82] M. R. Wick and W. B. Thompson. Reconstructive expert system explanation. *Artificial Intelligence*, 54(1-2):33–70, 1992. ISSN 00043702. doi: 10.1016/0004-3702(92)90087-E. URL https://www.sciencedirect.com/science/article/pii/000437029290087E.

[83] V. Perera and M. Veloso. Interpretability of a service robot: Enabling user questions and checkable answers. In *GCAI*, 2018.

[84] B. Krarup, M. Cashmore, D. Magazzeni, and T. Miller. Model-based contrastive explanations for explainable planning. In *ICAPS 2019 Workshop on Explainable AI Planning (XAIP)*, 2019.

[85] A. Tabrez, S. Agrawal, and B. Hayes. Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 249–257, March 2019. doi: 10.1109/HRI.2019.8673104.

[86] S. Sengupta, T. Chakraborti, S. Sreedharan, S. G. Vadlamudi, and S. Kambhampati. Radar - a proactive decision support system for human-in-the-loop planning. In *AAAI Fall Symposia*, 2017.

[87] N. Topin and M. Veloso. Generation of policy-level explanations for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2514–2521. /, 2019.

[88] K. Akash, K. Polson, T. Reid, and N. Jain. Improving Human-Machine Collaboration Through Transparency-based Feedback – Part I : Human Trust and Workload Model. *Elsevier*, 51(34):315–321, 2018. ISSN 24058963. doi: 10.1016/j.ifacol.2019.01.028. URL https://www.sciencedirect.com/science/article/pii/S2405896319300308.

[89] K. Akash, T. Reid, and N. Jain. Improving Human-Machine Collaboration Through Transparency-based Feedback – Part II: Control Design and Synthesis. *Elsevier*, 51 (34):322–328, 2018. ISSN 24058963. doi: 10.1016/j.ifacol.2019.01.026. URL https://www.sciencedirect.com/science/article/pii/S240589631930028Xhttp://www.sciencedirect.com/science/article/pii/S240589631930028X.

[90] S. Ososky, T. Sanders, F. Jentsch, P. Hancock, and J. Y. C. Chen. Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In R. E. Karlsen, D. W. Gage, C. M. Shoemaker, and G. R. Gerhart, editors, *Unmanned Systems Technology XVI*, volume 9084, page 90840E, unknown, jun 2014. Spie. ISBN 9781628410211. doi: 10.1117/12.2050622. URL http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2050622.

[91] J. Y. C. Chen, M. J. Barnes, A. R. Selkowitz, and K. Stowers. Effects of Agent Transparency on human-autonomy teaming effectiveness. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1838–1843, Budapest, oct 2017. IEEE. ISBN 9781509018970. doi: 10.1109/SMC.2016.7844505. URL http://ieeexplore.ieee.org/document/7844505/.

[92] M. R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Error in Aviation*, 37:217–249, 2017. doi: 10.4324/9781315092898-13.

[93] C. Chao, M. Cakmak, and A. L. Thomaz. Transparent active learning for robots. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction (HRI)*, pages 317–324, Osaka, 2010. IEEE. ISBN 9781424448937. doi: 10.1109/HRI.2010.5453178. URL https://www.cc.gatech.edu/social-machines/papers/chao10{_}hri{_}transparent.pdf.

[94] M. W. Floyd and D. W. Aha. *Incorporating transparency during Trust-Guided behavior adaptation*, volume 9969 LNAI. Springer, Atlanta, 2016. ISBN 9783319470955. doi: 10.1007/978-3-319-47096-2_9.

[95] A. Kulkarni, S. G. Vadlamudi, Y. Zha, Y. Zhang, T. Chakraborti, and S. Kambhampati. Explicable planning as minimizing distance from expected behavior. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 4: 2075–2077, 2019. ISSN 15582914. URL https://dl.acm.org/doi/10.5555/3306127.3332015.

[96] H. Zhang and S. Liu. Design of autonomous navigation system based on affective cognitive learning and decision-making. In *2009 IEEE International Conference on Robotics and Biomimetics, ROBIO 2009*, pages 2491–2496, Guilin, 2009. IEEE, IEEE. ISBN 9781424447756. doi: 10.1109/ROBIO.2009.5420477.

[97] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *IJCAI*, 2017.

[98] S. Sreedharan, T. Chakraborti, and S. Kambhampati. Balancing explicability and explanation in human-aware planning. *ArXiv*, abs/1708.00543, 2017.

[99] Z. Gong and Y. Zhang. Behavior explanation as intention signaling in human-robot teaming. In *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1005–1011, 08 2018. doi: 10.1109/ROMAN.2018.8525675.

[100] A. Tabrez, S. Agrawal, and B. Hayes. Explanation-based reward coaching to improve human performance via reinforcement learning. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 249–257, 2019.

[101] K. Baraka and M. M. Veloso. Mobile Service Robot State Revealing Through Expressive Lights: Formalism, Design, and Evaluation. *International Journal of Social Robotics*, 10 (1):65–92, jan 2018. ISSN 18754805. doi: 10.1007/s12369-017-0431-x. URL https://doi.org/10.1007/s12369-017-0431-x.

[102] M. Kwon, S. H. Huang, and A. D. Dragan. Expressing Robot Incapability. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 87–95. IEEE Computer Society, feb 2018. ISBN 9781450349536. doi: 10.1145/3171221.3171276.

[103] X. Huang, C. Du, Y. Peng, X. Wang, and J. Liu. Goal-oriented action planning in partially observable stochastic domains. In *Proceedings - 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems, IEEE CCIS 2012*, volume 3, pages 1381–1385, Hangzhou, 2013. IEEE, IEEE. ISBN 9781467318556. doi: 10.1109/CCIS.2012.6664612.

[104] K. E. Schaefer, R. W. Brewer, J. Putney, E. Mottern, J. Barghout, and E. R. Straub. Relinquishing manual control: Collaboration requires the capability to understand robot intent. In *2016 International Conference on Collaboration Technologies and Systems (CTS)*, pages 359–366, Oct 2016. doi: 10.1109/CTS.2016.0071.

[105] E. C. Grigore, A. Roncone, O. Mangin, and B. Scassellati. Preference-Based Assistance Prediction for Human-Robot Collaboration Tasks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4441–4448, Madrid, dec 2018. IEEE. doi: 10.1109/IROS.2018.8593716.

[106] B. Brown and E. Laurier. The trouble with autopilots: Assisted and autonomous driving on the social road. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, volume 2017-May, pages 416–429, Denver, 2017. ACM. doi: 10.1145/3025453.3025462. URL `http://dx.doi.org/10.1145/3025453.3025462`.

[107] A. William Evans, M. Marge, E. Stump, G. Warnell, J. Conroy, D. Summers-Stay, and D. Baran. The future of human robot teams in the army: Factors affecting a model of human-system dialogue towards greater team collaboration. In *Advances in Human Factors in Robots and Unmanned Systems*, volume 499, pages 197–210, Walt Disney World, 2017. Springer. ISBN 9783319419589. doi: 10.1007/978-3-319-41959-6_17.

[108] I. Lutkebohle, J. Peltason, L. Schillingmann, B. Wrede, S. Wachsmuth, C. Elbrechter, and R. Haschke. The curious robot - structuring interactive robot learning. In *2009 IEEE International Conference on Robotics and Automation*, pages 4156–4162, May 2009. doi: 10.1109/ROBOT.2009.5152521.

[109] M. Lamb, R. Mayr, T. Lorenz, A. A. Minai, and M. J. Richardson. The Paths We Pick Together: A Behavioral Dynamics Algorithm for an HRI Pick-and-Place Task. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 165–

166, Chicago, mar 2018. IEEE. ISBN 9781450356152. doi: 10.1145/3173386.3177022. URL https://dl.acm.org/citation.cfm?id=3177022.

[110] P. Legg, J. Smith, and A. Downing. Visual analytics for collaborative human-machine confidence in human-centric active learning tasks. *Human-centric Computing and Information Sciences*, 9(5):1 – 25, dec 2019. ISSN 21921962. doi: 10.1186/s13673-019-0167-8.

[111] A. L. Thomaz and C. Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6):716–737, 2008. doi: 10.1016/j.artint.2007.09.009. URL https://www.cc.gatech.edu/{~}athomaz/papers/ThomazBreazeal-AIJ.pdf.

[112] M. Arnold, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, K. R. Varshney, R. K. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, and A. Olteanu. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4-5):6:1 – 6:13, 2019. doi: 10.1147/JRD.2019.2942288.

[113] N. Wang, D. V. Pynadath, and S. G. Hill. Trust calibration within a human-robot team: Comparing automatically generated explanations. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 109–116, 2016.

[114] A. Roncone, O. Mangin, and B. Scassellati. Transparent role assignment and task allocation in human robot collaboration. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1014–1021, May 2017. doi: 10.1109/ICRA.2017.7989122.

[115] A. Sciutti, L. Patanè, F. Nori, and G. Sandini. Understanding object weight from human and humanoid lifting actions. *IEEE Transactions on Autonomous Mental Development*, 6 (2):80–92, 2014. ISSN 19430604. doi: 10.1109/TAMD.2014.2312399. URL https://ieeexplore.ieee.org/abstract/document/6776437/.

[116] A. Poulsen, O. K. Burmeister, and D. Tien. Care Robot Transparency Isn't Enough for Trust. In *2018 IEEE Region 10 Symposium, Tensymp 2018*, pages 293–297, Sydney, jul 2018. IEEE. ISBN 9781538669891. doi: 10.1109/TENCONSpring.2018.8692047. URL https://ieeexplore.ieee.org/document/8692047/.

[117] T. Chakraborti, A. Kulkarni, S. Sreedharan, D. E. Smith, and S. Kambhampati. Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In *ICAPS*, 2019.

[118] N. M. Seel, editor. *Imitation Learning*, pages 1494–1494. Springer US, Boston, MA, 2012. ISBN 978-1-4419-1428-6. doi: 10.1007/978-1-4419-1428-6_4288. URL https://doi.org/10.1007/978-1-4419-1428-6_4288.

[119] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters. An algorithmic perspective on imitation learning. *Found. Trends Robotics*, 7:1–179, 2018.

[120] P. Abbeel and A. Y. Ng. *Inverse Reinforcement Learning*, pages 554–558. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_417. URL https://doi.org/10.1007/978-0-387-30164-8_417.

[121] D. A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *NIPS*, 1988.

[122] B. D. Ziebart, N. D. Ratliff, G. Gallagher, C. Mertz, K. M. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. S. Srinivasa. Planning-based prediction for pedestrians. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3931–3936, 2009.

[123] M. E. Taylor and P. Stone. Behavior transfer for value-function-based reinforcement learning. In *AAMAS '05*, 2005.

[124] P. Abbeel, A. Coates, M. Quigley, and A. Ng. An application of reinforcement learning to aerobatic helicopter flight. In *NIPS*, 2006.

[125] A. Billard and D. H. Grollman. Robot learning by demonstration. *Scholarpedia*, 8:3824, 2013.

[126] R. E. Kálmán. When is a linear control system optimal. *Journal of Basic Engineering*, 86: 51–60, 1964.

[127] S. J. Russell. Learning agents for uncertain environments (extended abstract). In *COLT' 98*, 1998.

[128] D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3:88–97, 1991.

[129] F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. *ArXiv*, abs/1805.01954, 2018.

[130] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *IJCAI*, 2007.

[131] A. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In *ICML*, 2000.

[132] A. G. Barto, R. S. Sutton, and C. W. Anderson. Looking back on the actor-critic architecture. *IEEE Trans. Syst. Man Cybern. Syst.*, 51(1):40–50, 2021. doi: 10.1109/TSMC.2020.3041775. URL https://doi.org/10.1109/TSMC.2020.3041775.

[133] P. Abbeel and A. Ng. Apprenticeship learning via inverse reinforcement learning. *Proceedings of the twenty-first international conference on Machine learning*, 2004.

[134] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.

[135] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008.

[136] M. Lopes, F. S. Melo, and L. Montesano. Active learning for reward estimation in inverse reinforcement learning. In *ECML/PKDD*, 2009.

[137] M. L. Littman. Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 521:445–451, 2015.

[138] S. Ross and D. Bagnell. Efficient reductions for imitation learning. In Y. W. Teh and M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 661–668, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/ross10a.html.

[139] U. Syed and R. E. Schapire. A reduction from apprenticeship learning to classification. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/5c572eca050594c7bc3c36e7e8ab9550-Paper.pdf.

[140] S. Ross, G. J. Gordon, and J. A. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

[141] H. Daumé, J. Langford, and D. Marcu. Search-based structured prediction. *Mach. Learn.*, 75(3):297–325, jun 2009. ISSN 0885-6125. doi: 10.1007/s10994-009-5106-x. URL https://doi.org/10.1007/s10994-009-5106-x.

[142] J. MacGlashan, M. K. Ho, R. T. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman. Interactive learning from policy-dependent human feedback. *ArXiv*, abs/1701.06049, 2017.

[143] T. Ellman. Explanation-based learning: a survey of programs and perspectives. *ACM Comput. Surv.*, 21:163–221, 1989. doi: 10.1145/66443.66445. URL https://www.semanticscholar.org/paper/4c3a2306d23c5a2823f282c93a10c7625021d369.

[144] C. H. Lewis. Why and how to learn why: Analysis-based generalization of procedures. *Cogn. Sci.*, 12:211–256, 1988.

[145] J. Shavlik and G. DeJong. Acquiring general iterative concepts by reformulating explanations observed examples. *Machine Learning*, pages 302–350, 1990. doi: 10.1016/B978-0-08-051055-2.50018-3. URL https://www.semanticscholar.org/paper/45c058a98e8111bd6d140f3072dadb2b47c74370.

[146] J. Wusteman. Explanation-based learning: A survey. *Artificial Intelligence Review*, 6:243–262, 1992. doi: 10.1007/BF00155763. URL https://www.semanticscholar.org/paper/2cb09607b16a7ab2d560f405f428cdfef664da6f.

[147] J. Laird, P. Rosenbloom, and A. Newell. Chunking in soar: The anatomy of a general learning mechanism. *Machine Learning*, 1:11–46, 1986. doi: 10.1023/A:1022639103969. URL https://www.semanticscholar.org/paper/836f2cdc78e80c625e683008b9d6a2d20b9e192f.

[148] R. Keller. Defining operationality for explanation-based learning. *Artif. Intell.*, 35:227–241, 1987. doi: 10.1016/0004-3702(88)90013-6. URL https://www.semanticscholar.org/paper/4ef7369b076dc4c10a18de6f70f143d2eab2372d.

[149] R. P. Hall. Computational approaches to analogical reasoning: A comparative analysis. *Artificial Intelligence*, 39(1):39–120, 1989. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(89)90003-9. URL https://www.sciencedirect.com/science/article/pii/0004370289900039.

[150] A. Segre. On the operationality/generality trade-off in explanation-based learning. *IJCAI*, 1987. URL https://www.semanticscholar.org/paper/e6cced2ac26832ce41badc5b8338ca833ae14453.

[151] S. Kambhampati, S. Katukam, and Y. Qu. Failure driven dynamic search control for partial order planners: An explanation based approach. *Artif. Intell.*, 88:253–315, 1996. doi: 10.1016/S0004-3702(96)00005-7. URL https://semanticscholar.org/paper/99b8179c17cc560fb894fc9a8b5600a3c053094f.

[152] M. Rayner. Applying explanation-based generalization to natural-language processing. In *FGCS*, 1988.

[153] R. Mooney and G. DeJong. Learning schemata for natural language processing. *IJCAI*, 1985. URL https://www.semanticscholar.org/paper/462b0ebe93eee8b7225d6ce6b0df3d94385ab09f.

[154] T. M. Mitchell and S. Thrun. Explanation-based neural network learning for robot control. In *NIPS*, 1992.

[155] M. Babes-Vroman, J. MacGlashan, R. Gao, K. Winner, R. Adjogah, M. desJardins, M. S. Littman, and S. Muresan. Learning to interpret natural language instructions. In -, 2012.

[156] T. R. Sumers, M. K. Ho, and T. L. Griffiths. Show or tell? demonstration is more robust to changes in shared perception than explanation. In -, 2020.

[157] M. S. Lee, H. Admoni, and R. Simmons. Machine teaching for human inverse reinforcement learning. *Frontiers in Robotics and AI*, 8, 2021.

[158] C. Breazeal, G. Hoffman, and A. L. Thomaz. Teaching and working with robots as a collaboration. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, pages 1030–1037, 2004.

[159] P. Dillenbourg, M. Baker, A. Blaye, and C. O'Malley. The evolution of research on collaborative learning. In *Towards an interdisciplinary learning science. Oxford: Elsevier.*, 1996.

[160] L. Fiala and S. Suetens. Transparency and cooperation in repeated dilemma games: a meta study. *Experimental economics*, 20(4):755–771, 2017.

[161] D. Fudenberg and E. Maskin. *The Folk Theorem in Repeated Games with Discounting or With Incomplete Information*. The Econometric Society, 2009. doi: 10.1142/9789812818478_0011.

[162] D. Davis, O. Korenok, and R. Reilly. Cooperation without coordination: Signaling, types and tacit collusion in laboratory oligopolies. *Experimental economics*, 13(1):45–65, 2010.

[163] A. Tabrez, M. B. Luebbers, and B. Hayes. A survey of mental modeling techniques in human–robot teaming. In *Springer Curr Robot Rep*, 2020.

[164] M. Faria, F. S. Melo, and A. Paiva. Understanding robots: Making robots more legible in multi-party interactions. *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 1031–1036, 2021.

[165] K. Fischer, H. M. Weigelin, and L. Bodenhagen. Increasing trust in human-robot medical interactions: Effects of transparency and adaptability. *Paladyn*, 9(1):95–109, 2018. ISSN 20814836. doi: 10.1515/pjbr-2018-0007. URL https://doi.org/10.1515/pjbr-2018-0007.

[166] L. Perlmutter, E. M. Kernfeld, and M. Cakmak. Situated Language Understanding with Human-like and Visualization-Based Transparency. In *Robotics: Science and Systems*, volume 12, pages 40–50, Michigan, 2016. IEEE. ISBN 9780992374723. doi: 10.15607/rss.2016.xii.040. URL http://wiki.ros.org/openni.

[167] M. W. Floyd and D. W. Aha. Using explanations to provide transparency during trust-guided behavior adaptation 1. *AI Communications*, 30(3-4):281–294, 2017. ISSN 09217126. doi: 10.3233/AIC-170733.

[168] N. Wang, D. V. Pynadath, and S. G. Hill. The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, AAMAS '16, pages 997–1005, Richland, SC, 2016. ACM. ISBN 978-1-4503-4239-1. URL https://dl.acm.org/citation.cfm?id=2937071.

[169] N. Wang, D. V. Pynadath, E. Rovira, M. Barnes, and S. G. Hill. Is it my looks? or something i said? the impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In *PERSUASIVE*, 2018.

[170] A. Zhou, D. Hadfield-Menell, A. Nagabandi, and A. D. Dragan. Expressive Robot Motion Timing. In *ACM/IEEE International Conference on Human-Robot Interaction*, volume Part F1271, pages 22–31, New York, New York, USA, mar 2017. IEEE. doi: 10.1145/2909824.3020221. URL http://dl.acm.org/citation.cfm?doid=2909824.3020221http://arxiv.org/abs/1802.01536http://dx.doi.org/10.1145/2909824.3020221.

[171] M. W. Boyce, J. Y. C. Chen, A. R. Selkowitz, and S. G. Lakhmani. Effects of Agent Transparency on Operator Trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, volume 02-05-Marc of *HRI'15 Extended Abstracts*, pages 179–180, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-

3318-4. doi: 10.1145/2701973.2702059. URL http://dl.acm.org/citation.cfm?doid=2701973.2702059.

[172] M. Khoramshahi and A. Billard. A dynamical system approach to task-adaptation in physical human–robot interaction. *Autonomous Robots*, 43(4):927–946, apr 2019. ISSN 15737527. doi: 10.1007/s10514-018-9764-z.

[173] M. S. Lee. Self-Assessing and Communicating Manipulation Proficiency Through Active Uncertainty Characterization. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, volume 2019-March, pages 724–726, Daegu, mar 2019. IEEE. ISBN 9781538685556. doi: 10.1109/HRI.2019.8673083.

[174] K. Baraka, S. Rosenthal, and M. M. Veloso. Enhancing human understanding of a mobile robot ' s state and actions using expressive lights Enhancing Human Understanding of a Mobile Robot's State and Actions using Expressive Lights. In *ieeexplore.ieee.org*. IEEE, 2018. doi: 10.1109/ROMAN.2016.7745187.

[175] S. Sheikholeslami, J. W. Hart, W. P. Chan, C. P. Quintero, and E. A. Croft. Prediction and Production of Human Reaching Trajectories for Human-Robot Interaction. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 321–322, Chicago, mar 2018. IEEE. ISBN 9781450356152. doi: 10.1145/3173386.3176924.

[176] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. E. Rybski. Gracefully mitigating breakdowns in robotic services. *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 203–210, 2010.

[177] T. Chakraborti, S. Sreedharan, A. Kulkarni, and S. Kambhampati. Projection-Aware Task Planning and Execution for Human-in-the-Loop Operation of Robots in a Mixed-Reality Workspace. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4476–4482, Madrid, 2018. IEEE. ISBN 9781538680940. doi: 10.1109/IROS.2018.8593830. URL https://ieeexplore.ieee.org/abstract/document/8593830/.

[178] J. Y. C. Chen, S. G. Lakhmani, K. Stowers, A. R. Selkowitz, J. L. Wright, and M. Barnes. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19(3):259–282, 2018. doi: 10.1080/1463922X.2017.1315750. URL https://doi.org/10.1080/1463922X.2017.1315750.

[179] A. Rosenfeld and A. Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6):673–705, nov 2019. ISSN 15737454. doi: 10.1007/s10458-019-09408-y.

[180] J. Zelmer. Linear public goods experiments: A meta-analysis. *Experimental Economics*, 6: 299–310, 2001.

[181] M. N. Burton-Chellew, C. E. Mouden, and S. West. Conditional cooperation and confusion in public-goods experiments. *Proceedings of the National Academy of Sciences*, 113:1291 – 1296, 2016.

[182] J. Y. Chen and M. J. Barnes. Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, 44(1):13–29, 2014.

[183] J. Y. Chen and M. J. Barnes. Agent transparency for human-agent teaming effectiveness. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1381–1385. IEEE, 2015.

[184] J. E. Mercado, M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, and K. Procci. Intelligent agent transparency in human–agent teaming for multi-uxv management. *Human factors*, 58(3):401–415, 2016.

[185] T. Helldin. Transparency for future semi-automated systems. *PhD diss., Orebro University*, 2014.

[186] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250. ACM, 2000.

[187] F. Correia, S. F. Mascarenhas, S. Gomes, P. Arriaga, I. Leite, R. Prada, F. S. Melo, and A. Paiva. Exploring prosociality in human-robot teams. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 143–151. IEEE, 2019.

[188] K. Allen and R. Bergin. Exploring trust, group satisfaction, and performance in geographically dispersed and co-located university technology commercialization teams. In *In Proceedings of the NCIIA 8th Annual Meeting: Education that Works*, pages 18–20, 2004.

[189] C. W. Leach, M. Van Zomeren, S. Zebel, M. L. Vliek, S. F. Pennekamp, B. Doosje, J. W. Ouwerkerk, and R. Spears. Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. *Journal of personality and social psychology*, 95(1):144, 2008.

[190] C. Bartneck, D. Kulic, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1:71–81, 01 2008. doi: 10.1007/s12369-008-0001-3.

[191] U. Segal and J. Sobel. Tit for tat: Foundations of preferences for reciprocity in strategic settings. *Journal of Economic Theory*, 136(1):197–216, 2007. URL https://EconPapers.repec.org/RePEc:eee:jetheo:v:136:y:2007:i:1:p:197-216.

[192] R. Axelrod. On six advances in cooperation theory. *Analyse Kritik*, 22:130–151, 08 2000. doi: 10.1515/auk-2000-0107.

[193] T. Lombrozo and N. Z. Gwynne. Explanation and inference: mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, 8, 2014.

[194] F. Keil. Explanation and understanding. *Annual review of psychology*, 57:227–54, 2006.

[195] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009. ISBN 0136042597, 9780136042594.

[196] D. Kahneman and A. Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291, March 1979.

[197] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[198] A. Bandura, W. Freeman, and R. Lightsey. Self-efficacy: The exercise of control, 1999.

[199] C. Pastorelli, G. Caprara, C. Barbaranelli, J. Rola, S. Rózsa, and A. Bandura. The structure of children's perceived self-efficacy: A cross-national study. *European Journal of Psychological Assessment*, 17:87–97, 05 2001. doi: 10.1027//1015-5759.17.2.87.

[200] A. Billard, S. Calinon, R. Dillmann, and S. Schaal. Robot programming by demonstration. In *Springer Handbook of Robotics*, 2008.

[201] J. F. Montgomery and G. A. Bekey. Learning helicopter control through "teaching by showing". *Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No.98CH36171)*, 4:3647–3652 vol.4, 1998.

[202] Ny Vasila and Azzurra Ruggeri and Tania Lombrozo. When and how children use explanations to guide generalizations. *Cognitive Development*, 61, 2021.

[203] Michael M Chouinard. Children's questions: a mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, 72:vii–126, 2007.

[204] J. Choi and K. Kim. MAP inference for bayesian inverse reinforcement learning. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1989–1997, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/3a15c7d0bbe60300a39f76f8a5ba6896-Abstract.html.

[205] A. J. Chan and M. van der Schaar. Scalable bayesian inverse reinforcement learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=4qR3coiNaIv.

[206] D. Hadfield-Menell, S. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. In *NIPS*, 2016.

[207] D. S. Brown, W. Goo, P. Nagarajan, and S. Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 783–792. PMLR, 2019. URL http://proceedings.mlr.press/v97/brown19a.html.

[208] A. Theodorou, R. H. Wortham, and J. J. Bryson. Why is my Robot Behaving Like That? Designing Transparency for Real Time Inspection of Autonomous Robots. In *AISB Workshop on Principles of Robotics*, page 4, Sheffield, 2016. unknown. URL http://opus.bath.ac.uk/49713/.

[209] M. C. Vroman. *Maximum Likelihood Inverse Reinforcement Learning*. University of New Brunswick Rutgers, The State University of New Jersey, 2014.

[210] O. Z. Khan, P. Poupart, and J. P. Black. Minimal sufficient explanations for factored markov decision processes. In *Proceedings of the Nineteenth International Conference on International Conference on Automated Planning and Scheduling*, ICAPS'09, page 194–200. AAAI Press, 2009. ISBN 9781577354062.

[211] J. Zhang and E. Bareinboim. Fairness in decision-making — the causal explanation formula. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11564.

[212] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 1995.

[213] R. BELLMAN. A markovian decision process. *Journal of Mathematics and Mechanics*, 6 (5):679–684, 1957. ISSN 00959057, 19435274. URL http://www.jstor.org/stable/24900506.

[214] T. Sumers, M. K. Ho, R. D. Hawkins, and T. Griffiths. Show or tell? teaching with language outperforms demonstration but only when context is shared. *PsyArXiv*, 2021.

[215] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan. Towards a science of human-ai decision making: A survey of empirical studies. *ArXiv*, abs/2112.11471, 2021.

[216] B. Hayes. Supportive behaviors for human-robot teaming. In *PhD Thesis*, 2016.

[217] B. N. Patel, L. B. Rosenberg, G. Willcox, D. Baltaxe, M. Lyons, J. A. Irvin, P. Rajpurkar, T. J. Amrhein, R. Gupta, S. S. Halabi, C. Langlotz, E. Lo, J. G. Mammarappallil, A. J. Mariano, G. Riley, J. Seekins, L. Shen, E. Zucker, and M. P. Lungren. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digital Medicine*, 2, 2019.

[218] Q. Yang, A. Steinfeld, and J. Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[219] K. van den Bosch, T. A. J. Schoonderwoerd, R. A. M. Blankendaal, and M. A. Neerincx. Six challenges for human-ai co-learning. In *HCI*, 2019.

[220] D. Wang, E. Churchill, P. Maes, X. Fan, B. Shneiderman, Y. Shi, and Q. Wang. From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–6, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368193. doi: 10.1145/3334480.3381069. URL https://doi.org/10.1145/3334480.3381069.

[221] J. Y. Halpern and R. Pass. Game theory with translucent players. *International Journal of Game Theory*, 47:949–976, 2013.

[222] P. Lange, W. Otten, E. M. N. De Bruin, and J. Joireman. Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of personality and social psychology*, 73:733–46, 11 1997. doi: 10.1037//0022-3514.73.4.733.

[223] J. J. Williams, T. Lombrozo, and B. Rehder. The hazards of explanation: overgeneralization in the face of exceptions. *Journal of experimental psychology. General*, 142 4:1006–14, 2013.

[224] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 676–682. IEEE, 2017.

[225] T. Hitron, Y. Orlev, I. Wald, A. Shamir, H. Erel, and O. Zuckerman. Can children understand machine learning concepts? the effect of uncovering black boxes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[226] R. Stower and A. Kappas. "oh no, my instructions were wrong!" an exploratory pilot towards children's trust in social robots. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 641–646, 2020. doi: 10.1109/RO-MAN47096.2020.9223495.

[227] E. Yadollahi, W. Johal, A. Paiva, and P. Dillenbourg. When deictic gestures in a robot can harm child-robot collaboration. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*, pages 195–206. ACM, 2018.

[228] S. Lemaignan and P. Dillenbourg. Mutual modelling in robotics: Inspirations for the next steps. *ACM/IEEE International Conference on Human-Robot Interaction*, 2015:303–310, 03 2015. doi: 10.1145/2696454.2696493.

[229] A. Ramachandran, C.-M. Huang, and B. Scassellati. Toward effective robot–child tutoring. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 9:1 – 23, 2019.

[230] C. Conati, K. Porayska-Pomsta, and M. Mavrikis. Ai in education needs interpretable machine learning: Lessons from open learner modelling, 2018.

[231] L. Buesing, T. Weber, Y. Zwols, S. Racanière, A. Guez, J.-B. Lespiau, and N. M. O. Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *ArXiv*, abs/1811.06272, 2019.

[232] J. Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, 2019.

[233] D. Freedman. From association to causation: some remarks on the history of statistics. *Statistical Science*, 14:243–258, 1999.

[234] D. Hume. A treatise of human nature. 1970.

[235] S. Glennan. Mechanisms and the nature of causation. *Erkenntnis*, 44:49–71, 1996.

[236] N. Vélez and H. Gweon. Learning from other minds: an optimistic critique of reinforcement learning models of social learning. *Current Opinion in Behavioral Sciences*, 38:110–115, 2021.

[237] A. Rosenfeld. Better metrics for evaluating explainable artificial intelligence. In *AAMAS*, 2021.

[238] J. Henrich. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princenton University Press, 2015.

Appendix A

List of Publications

Journal publications

- Silvia Tulli, Francisco S. Melo, Ana Paiva, Mohamed Chetouani (2022). Learning from Explanations with Maximum Likelihood Inverse Reinforcement Learning. Submitted to Neural Computing and Applications, Topical Collection on Human-aligned Reinforcement Learning for Autonomous Agents and Robots [under review]

- Silvia Tulli*, Sebastian Wallkötter*, Ginevra Castellano, Ana Paiva, Mohamed Chetouani (2021). Explainable Embodied Agents Through Social Cues: A Review. ACM Transactions on Human-Robot Interaction, THRI Special issue in on Explainable Robotic Systems *equal contribution

Preprint

- Adrien Bennetot, Ivan Donadello, Ayoub El Qadi, Mauro Dragoni, Thomas Frossard, Maria Trocan, Raja Chatila, Benedikt Wagner, Anna Saranti, Andreas Holzinger, Silvia Tulli, Artur d'Avila Garcez, Natalia Díaz-Rodríguez (2021). A Practical Tutorial on Explainable AI Techniques [Arxiv]

Conference Papers

- Silvia Tulli, Marta Couto, Miguel Vasco, Elmira Yadollahi, Francisco S. Melo, Ana Paiva (2020). Explainable Agency by Revealing Suboptimality in Child-Robot Learning Scenarios, 12th International Conference on Social Robotics. Springer post-proceedings Lecture Notes in Artificial Intelligence, Best Student Paper Award

- Silvia Tulli, Diego Agustin Ambrossio, Amro Najjar, and Francisco Javier Rodríguez Lera (2019). Great Expectations & Aborted Business Initiatives: The Paradox of Social Robot Between Research and Industry, Proceedings of the Reference AI & ML Conference for Belgium, Netherlands & Luxemburg

- Filipa Correia, Samuel Mascarenhas, Samuel Gomes, Silvia Tulli, Fernando P Santos, Francisco C Santos, Rui Prada, Francisco S Melo, Ana Paiva (2019). For The Record-A Public Goods Game For Exploring Human-Robot Collaboration (Demo), Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems

Doctoral Consortium

- Silvia Tulli (2020). Explainability in Autonomous Pedagogical Agents. Proceedings of the AAAI Conference on Artificial Intelligence

Workshop Papers

- Silvia Tulli, Sebastian Wallkötter, Ana Paiva, Francisco S. Melo, Mohamed Chetouani (2020). Learning from Explanations and Demonstrations: A Pilot Study, NL4XAI2020 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, part of the 13th International Conference on Natural Language Generation (INLG2020)
- Silvia Tulli, Filipa Correia, Samuel Mascarenhas, Samuel Gomes, Francisco S. Melo, Ana Paiva (2019). Effects of Agents' Transparency on Teamwork. 1st International Workshop on EXplainable TRansparent Autonomous Agents and Multi-Agent Systems at International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2019. Springer post-proceedings Lecture Notes in Artificial Intelligence
- Patrícia Alves-Oliveira, Silvia Tulli, Philipp Wilken, Ramona Merhej, João Gandum, and Ana Paiva (2019). Sparking Creativity with Robots: A Design Perspective. Robots for Social Good: Exploring Critical Design for HRI Workshop at 14th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2019

Reports

- Reports of the Workshops Held at the 2022 AAAI Conference on Artificial Intelligence
- Reports of the Workshops Held at the 2021 AAAI Conference on Artificial Intelligence