UNIVERSITY OF GENOVA

PHD PROGRAM IN BIOENGINEERING AND ROBOTICS

XAI in HRI: A Journey to the Centre of the Explainability

by

Marco Matarese

Thesis submitted for the degree of *Doctor of Philosophy* (36° cycle)

December 2023

Prof. Giulio Sandini	Supervisor
Dr. Alessandra Sciutti	Supervisor
Dr. Francesco Rea	Supervisor
Prof. Paolo Massobrio	Head of the PhD program

Thesis Jury:

Prof. Tim Miller, *University of Queensland*	External examiner
Prof. Marta Romeo, *Heriot-Watt University*	External examiner
Prof. Gualtiero Volpe, *University of Genoa*	Internal examiner

Dibris

Department of Informatics, Bioengineering, Robotics and Systems Engineering

To those who are here
to those who are gone
to those who will come.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Marco Matarese
March 2024

Acknowledgements

I want to thank my supervisors, Prof. Giulio Sandini, Dr. Alessandra Sciutti, and Dr. Francesco Rea, for their support and inspiration during these years. You are witnesses of passion, determination, and goodness. I feel blessed to have you guiding my academic path.

A big hug goes to Prof. Katharina Rohlfing and her team at the University of Paderborn and the University of Bielefeld, who warmly welcomed me during my period abroad. Collaboration and friendship are stronger than distance and cultural differences: I experienced it and will always try to bring with me this valuable lesson.

Last but not least, I would love to thank Prof. Guglielmo Tamburrini for taking my breath away during our university classes, and inspiring me to try to give back to the world that spark of beauty and passion that I saw every day in his eyes.

Abstract

We are witnessing the global spread of artificial intelligence (AI) technology in people's everyday lives. With the advent of the AI surge, the demand for explainable AI (XAI) techniques arose due to the growing intricacy of AI models. Users sought to comprehend the reasoning behind the decisions made by these models, a necessity that became more pressing as an expanding number of AI-driven robots interacted with people in real-world scenarios.

Recent years have witnessed the XAI community recognizing the imperative of leveraging the social dimensions of the explanation process. Drawing insights from psychology and cognitive sciences, researchers are increasingly reframing explainability as a social problem. This conceptual shift is exemplified by explainable robots capable of using the common ground established with human partners during the explanation generation process.

Social autonomous robots can trigger neural and social mechanisms in people similar to those happening during interactions between humans. Given the ease with which individuals attribute intentions, beliefs, and even second-order theory of mind capabilities to robots, the human-robot interaction (HRI) community is increasingly incorporating these mechanisms into explanatory interchanges.

This thesis introduces a theoretical framework for XAI in HRI that leverages the social-dialogical nature of explanations. The framework models the explanation process as a dialogue between robots and human partners. Differently form existing approaches, our framework emphasizes the influence of the human-robot common ground and interaction history on the generation of explanations.

Subsequently, the framework's philosophy is implemented in an HRI collaborative decision-making scenario. The focus is on exploring how explanations based on shared human-robot experiences impact individuals' decision-making and the role of their personality traits in this context. Results showed that a social robot that justifies its suggestions with explanations exploiting its common ground with the human partner is more persuasive than classical explanations, especially for less skilled participants. Moreover, participants' personality traits significantly impacted their decision-making and interaction with the robot.

Finally, to assess the effectiveness of such explanations compared to classical ones, an evaluation task is designed to measure the informativeness of XAI systems for non-expert users. This task is instantiated in various domains such as human-computer interaction (HCI), HRI, and self-learning, examining how different types of explanations and artificial explainable agents influence people's learning of new tasks. Results showed that expert explainable agents influenced participants' learning, limiting them from adequately exploring the learning environment, as participants who learned alone did.

Through this thesis, we advanced existing literature on collaborative decision-making with both HCI and HRI domains. Employing methodologies derived from HRI, we compared classical XAI techniques with explanation approaches that leverage on the human-robot common ground. Our investigation highlighted how these latter approaches improve robot's persuasiveness, particularly in social collaborative contexts. Additionally, we conceptualized and developed an assessment task to measure the quality of explanations. Our findings did not highlight differences between classical and partner-aware explanation methodologies. Nevertheless, results brought to light the influence that both robotic and artificial agents have on people's learning, limiting their exploration strategies.

Table of contents

List of figures

List of tables

Part I

Introduction

Chapter 1

Mission statement

"Courage, dear heart."
Aslan - C.S. Lewis, The Voyage of the Dawn Treader

1.1 Motivation

This thesis seeks to advance our comprehension of the intricate dynamics between humans and explainable robots within collaborative decision-making tasks. Collaboration stands as a pivotal factor in the evolution of humanity and the development of societies, facilitated by intricate mechanisms such as behaviour anticipation and intention attribution. The human-robot interaction (HRI) field has established that people rely on similar mechanisms when engaging with artificial agents, with a heightened emphasis on intention attribution when robots provide rationales for their decisions (Ahmad et al., 2017; Millecamp et al., 2019).

Motivated by these considerations, we investigated the interactive dynamics between humans and explainable robots. Specifically, we explored the design principles for explainable robots capable of leveraging collaborative mechanisms and examined their influence on humans during collaborative decision-making tasks. Similarly to what have been done in the human-computer interaction (HCI) field, we aim to guarantee people's control and freedom while interacting with explainable robots by preventing unacceptable human behaviors, such as over-reliance and automation (Nielsen, 2005).

Beyond mere information provision, explanation is recognised as a social act, entailing a dialogue between the explainer and the explainee (Miller, 2019; Rohlfing et al., 2020). Acknowledging the partner in this exchange offers advantages for both parties: the explainer can tailor explanations based on the explainee's understanding, while the explainee actively contributes to the route of the discourse. Despite our growing understanding of how people attribute mental states to explainable agents, there remains a limited exploration of the influence of explainable social robots in collaborative decision-making tasks (Anjomshoae et al., 2019; Wallkötter et al., 2021).

We envision a future where individuals collaborate with expert artificial agents. In the HCI field, the influence of explainable artificial intelligence (XAI) models is increasingly being studied (Bertrand et al., 2022a). Nonetheless, the influence of explainable social robots remains relatively unexplored. The social component and embodied capabilities of robots can prove pivotal in numerous scenarios, necessitating a closer examination of their impact.

To exploit the interactive facet of explanations within the XAI models, we advocate for personalised explanations using partner models: dynamic models to represent the human partner's internal states and current knowledge (Robrecht and Kopp, 2023). However, the efficacy of partner-aware XAI requires further scrutiny. While recent studies shed light on their influence during collaboration, the literature lacks robust benchmarks to conclusively determine the superiority of partner-aware approaches.

Measuring the quality of explanations poses significant challenges. XAI models' binding with the application context, the AI models used, and even the human user represents the main reasons why it is difficult to define how good explanations are. Recognising this, there is a pressing need in the XAI field to design objective and quantitative methods for comparing explanatory approaches.

In summary, the forthcoming challenge in social robotics lies in designing explainable robots that consider the individual needs of human explainees. Concurrently, the community must scrutinise the validity of partner-aware explanations and their impact on individuals, particularly in the context of human-robot collaboration, highlighting the urgency for comprehensive research in this domain.

1.1.1 Research questions

The established idea that robots exert influence on human behaviour, particularly in collaborative contexts, is well-acknowledged in the human-robot interaction (HRI) field (Sanders et al., 2014a). This work seeks to extend this understanding by delving into the impact of robots' explanation strategies, specifically those that take into account the human partner (partner-aware), on collaborative scenarios. For our experimentation, we chose the decision-making setting because it is particularly relevant for those fields (*e.g.*, healthcare and industrial robotics) where the human-robot cooperation is crucial to address complex problems and where decisions can have severe consequences. Here, we formulated the main research questions addressed throughout the thesis.

- **RQ1**. To what extent do partner-aware explanations influence individuals compared to classical explanations during collaborative decision-making tasks in HRI?
- **RQ2**. How do people's personality traits contribute to the dynamics of HRI when interacting with explainable robots during collaborative decision-making tasks?
- **RQ3**. Is there a significant increase in informativeness for non-expert users when exposed to partner-aware explanations as opposed to classical approaches?
- **RQ4**. What is the impact of expert explainable artificial agents on non-expert users' learning processes when undertaking a new task?

1.2 Contribution to knowledge

To tackle RQ1, we designed a collaborative version of the *Connect 4* game, wherein participants played with the iCub robot to outperform the computer. Throughout the game, participants negotiated each move, and, contingent on the experimental condition, the robot provided justifications when its suggestions diverged from the participants' original ideas. Our comparison of two explanation strategies revealed that incorporating the HRI history significantly increased the likelihood of participants accepting iCub's suggestions compared to classical explanations. However, participants did not clearly prefer a specific XAI strategy, and both explanation types demonstrated comparable performance. Hence, partner-aware explanations demonstrate their power to influence individuals more compared to classical explanations.

Addressing RQ2 involved the same experiment, preceded by submitting a self-reported questionnaire to gather participants' personality traits. The focus of this research question centred on the dichotomy between "explainable vs. non-explainable" robots. Such a further investigated was motivated by the current interest in personalizing XAI exploiting users' personality traits, and the occurrence of inconsistent results between different works Abdulrahman et al. (2022); Conati et al. (2021); Kallina (2020). Our findings indicated that participants' negative agency and agreeableness played pivotal roles in shaping their inclination to follow suggestions from explainable robots. Our results suggest the occurrence of various human social mechanisms during the HRI scenario, such as one's tendency to rely a lot on robot's suggestions when they are familiar to the robot itself. Hence, participants' agreeableness and negative agency significantly influenced their tendency to rely on the explainable robot's suggestions.

We investigated RQ3 by designing an assessment task to measure the informativeness of explanations to non-expert users while learning a new task. Subsequently, we compared participants' results based on interactions with classical and partner-aware explainable artificial agents. We used two agents in this study: a virtual dialogical agent and the humanoid social robot iCub. The type of artificial agent did not yield substantial differences regarding participants' understanding of the task, and their performance remained comparable across the two groups. However, with the virtual agent, the partner-aware explanations induced a more decisive behaviour in participants, making them move faster than those who received classical explanations. While the robot's partner-aware explanations enhanced its persuasiveness, making participant more prone to accept the robot's suggestions. Hence, this peculiar setting showed how partner-aware explanations triggered different user behaviors when administered by different artificial agents.

To address RQ4, we replicated the same experiment without any agent to interact with, requiring participants to learn the task autonomously. We found that self-taught participants demonstrated performance comparable to their counterparts interacting with artificial agents. However, the former made more mistakes during the training than the latter because they were not guided by any agent; thus, their training resulted to be more risk-prone. Moreover, they felt free to deeply explore the environment via trial and errors, resulting in a more extensive comprehension of the task compared to those participants engaged with the artificial agents. Hence, the expert explainable agents influenced participants learning strategies; in particular, self-taught participants explored the learning environment more compared to the XAI-assisted ones.

1.3 Thesis structure

Chapter 2 provides an overview of the XAI literature. It illustrates the problem of explaining from both philosophical and psychological perspectives, exploring existing approaches addressing XAI in the context of collaborative decision-making. Subsequently, this chapter presents the current approaches to explainability in HRI and culminates in an examination of the challenges of evaluating the quality of explanations.

Chapter 3 introduces our theoretical framework for XAI in HRI. The framework emphasises the social-dialogical nature of explanations and speculates on both the interaction and explanation generation processes. It provides a starting point for a discussion about explainability from a social robotics viewpoint.

Chapter 4 is dedicated to our first user study, focusing on the impact of partner-aware explanations during collaborative decision-making tasks in HRI. Chapter 5 revisits the same experiment, focusing on the influence of people's personality traits during collaborative decision-making tasks with explainable robots.

Chapter 6 describes an assessment method to measure the goodness of an XAI system, intended as the extent of information it provides to non-expert users. Chapter 7 presents the results of our second user study, wherein we implemented the assessment task presented in the previous chapter. Here, we compare the influence of a virtual artificial agent and a social humanoid robot.

Chapter 8 introduces an alternative experimental condition where participants had no interaction with any agent. This chapter provides a comparative analysis, contrasting the results of this condition with those obtained with the explainable artificial agents.

Lastly, Chapter 9 offers a comprehensive summary of the undertaken work, discussing the progress made within this thesis and acknowledging its current limitations.

Chapter 2

Background and related works

"Some people risk their lives to conquer a mountain peak. No one, not even they themselves, can really explain why".
Michael Ende, The Neverending Story

2.1 Explainable artificial intelligence

In the scientific literature, we can find different definitions of explanation. Moreover, several related concepts to explanations differ between authors and articles. In this work, we take Lewis's definition of explanations. He defined the act to explain an event as *"to provide information about its causal history"*: someone who has such information tries to convey it to someone else (Lewis, 1987). In this meaning, we mostly referred to explanations as justification for an agent's action when this is asked a *why-question* (Dennett, 1989; Lewis, 1987; Malle, 2006). Throughout this work, we refer to the term "explainer" as the one who provides explanations and "explainee" as the one who asks for them (Miller, 2019).

In most cases, why-questions are *contrastive* (Lipton, 1990), meaning that the explainee does not want to be explained the occurrence of the event per se, but why it occurred in that case and not in some other counterfactual cases (Hilton, 1990). Moreover, several empirical pieces of research confirm the hypothesis that people tend to ask questions about events that they consider unexpected (Heider, 2013). For the sake of completeness, we should specify that explanations can also answer other types of questions (Van Bouwel and Weber, 2002). Nevertheless, contrastive why questions remain the most challenging and interesting ones in the context of this research.

Causality and counterfactuals are notions related to each other in the XAI field. In contrastive explanation, the counterfactuals are hypothetical outcomes different from the occurred event. Indeed, several causality models have been proposed in the years that focus on counterfactuals (Fair, 1979; Halpern and Pearl, 2005; Menzies and Price, 1993). In the present work, we used the formulation from Lipton (1990), who refers to the terms “fact” and “foil” to the occurred event and a counterfactual case, respectively. People often tend to leave implicit the foil in everyday interaction. Indeed, several works have proved that people are good at inferring implicit foils (Hesslow, 1988; Lombrozo, 2009). For example, implicit foil can emerge when dealing with abnormal situation: “why *fact* (rather than the normal case)?”.

Lombrozo (2006) argued that explanation is both a process and a product. On the one hand, it is a *cognitive* process because of the abductive inference for filling the gaps to determine the explanation of an event (Chin-Parker and Bradner, 2010). On the other hand, it is a *social* process because of the transfer of knowledge between who asks for explanations - the explainee - and who provides them - the explainer (Miller, 2019).

Concepts like intentions and attribution are at the basis of our research of sense on others’ behaviours in everyday life (Heider, 2013). Through the recognition of intentions for others’ actions, it is possible to explain the rationale behind them: people constantly attribute intentionality to others’ behaviours (Dennett, 1989; Malle, 2006), and they do so also with artificial agents when those display certain human-like traits (Heider and Simmel, 1944; Roselli et al., 2021).

Intentions are also a social construct, even if people use them as a matter of fact to understand others’ actions. Regardless of whether such actions are intentional, people use intentions as objective facts in their social interactions. However, intentions never walk alone. People use concepts as beliefs and desires to explain human actions (Kashima et al., 1998). For example, Malle (2011) highlighted that people try to relate desires, beliefs and intentions in making sense of behaviour.

The most efficient way to relate those concepts is through causality: *causal chains* are essential in explanation. We can define causal chains as paths of causes that bring together a set of events (Hilton et al., 2005). Hume et al. (2000), in his regularity theory of causation, claims that there is a cause between two types of events if events of the first type are always followed by events of the second. Hume’s definition is also about counterfactuals rather than just dependence: the cause of a particular event should be understood relative to a counterfactual case.

However, usually, people do not need to understand the complete causal chain of an event to have a good explanation. In their explanations, people tend to select only a few causes rather than provide an event’s entire causal history. People use several strategies to select causes (almost unconsciously): looking at the differences between fact and foil (Hesslow, 1988; Lipton, 1990), referring to unusual situations that occurred (Hilton and Slugoski, 1986), or preferring sufficient and necessary causes above the others (Lipton, 1990; Lombrozo, 2010).

We have already pointed out that explanations are a social product. This principle has been expressed with great precision by Hilton: *"the verb to explain is a three-place predicate: someone explains something to someone"* (Hilton, 1990). Thus, explanations take the form of social conversations. If this is true, in order to provide adequate explanations, the information provided by them should comply with the rules of conversation (*e.g.*, Grice's maxims (Grice, 1975)).

Moreover, the interactive aspect of XAI leads to interesting thoughts about the Theory of Mind (ToM): a set of mechanisms by which people make sense of others' behaviours, *e.g.*, in terms of beliefs, intentions and desires. The simplest one regards tracking the explanatory conversation already had: the explainer should at least keep track of what they have already explained (Miller, 2019). More generally speaking, using a ToM of the explainee to reply to their questions effectively leans on exploiting the explainee's beliefs and intentions during the explanation selection. In particular, the partner's beliefs and intentions can be considered during the explanations selection to disambiguate between different explanation goals. Thus, the explainer would consider the explainee's mental models to understand better how to close the gap that the requested explanation is meant to do (Sreedharan et al., 2021).

Early attempts at XAI lacked effectiveness for the end-users mainly because they were designed by and intended for computer scientists. This led to solutions developers intuitively thought were helpful for them without any objective measure of effectiveness (Vilone and Longo, 2020). At first, such XAI systems mainly aimed to help during the debugging of complex machine learning (ML) models. However, Doshi-Velez and Kim (2017) recognised a need for more rigour about what interpretability means also in ML and how to evaluate it. Nevertheless, in recent years, researchers have begun to focus on the design and evaluation of good explanations (Mohseni et al., 2018a).

Recently, Miller (2019) highlighted that XAI is a social problem, not only a computer science one. Hence, the idea that the XAI research should be inspired by how humans explain to each other has taken hold in recent years. In this regard, it is crucial that the explainer would consider the explainee during the explanation process (Kirsch, 2017). For example, perspective-taking is a crucial factor in explanations, especially when the explanation cites a false belief, which explains the action truly from the explainer's subjective perspective rather than in terms of objective reality (De Graaf and Malle, 2017). Other works tackled this problem through adaptation: Robrecht and Kopp (2023) implemented an explainer model that constructs and employs a partner model to tailor explanations during the interaction.

Furthermore, several works highlighted the necessity to consider the social perspective in explaining AI suggestions during collaborative decision-making (Ehsan et al., 2021). This seems particularly important when considering embodied agents (Wallkötter et al., 2021); in human-robot interaction (HRI), it is crucial to meet the humans' interactive standards of explanations during the design of social robots (Arnold et al., 2021).

More recently, Rohlfing et al. (2020) theorised that explainees and explainers are involved in the first person in the explanation generation process. They argued that both actively *co-construct* explanations. Such a co-constructive process regards not only the social/dialogical roles of the parties: both the explainer and explainee actively negotiate the *explanandum* and the *explanans*, which are the object of the explanation and the way to convey it. This is possible through continuous *monitoring* and *scaffolding* of the partner. Moreover, they adapt their behaviours to the partner through monitoring and scaffolding, enacting the social practice of explaining (Fisher et al., 2023). Also prior knowledge and common ground are crucial in structuring their explanatory goals and expectations (Malle, 2006).

2.2 Explanations in human-robot interaction

In the past, research objectives in explainability have been connected with the concepts of interpretability and transparency (Rohlfing et al., 2020). Interpretability has been defined as a description of the *"internals of a system"* (Gilpin et al., 2018) and often - but wrongly - equated to explainability (Ciatto et al., 2020; Miller, 2019). Explainable models aim at making understandable descriptions to the explainee because they are concerned with the explainee's understanding (Rohlfing et al., 2020). In this regard, explainable models are interpretable, but the opposite is not always true (Gilpin et al., 2018). However, recent attempts at robots' explainability showed an increase in the perceived transparency (Angelopoulos et al., 2024).

Moreover, explanations are social mainly because they take the form of conversations, as highlighted in the previous section (Grice, 1975; Hilton, 1990). Cawsey (1993) has already proposed the idea of integrating a user model into the explanatory system. Indeed, the EDGE user model consists of the knowledge the user has about a phenomenon (similarly to the BLAH system (Weiner, 1980)) and their level of expertise. However, integrating the partner model into the explanatory system still needs further investigation (Fisher et al., 2023).

Ciatto et al. (2020), according to the idea that explanations are a social process (Grice, 1975), proposed the idea of estimating the explainee's ToM (Premack and Woodruff, 1978) to understand their interpretation of the environment better. For this reason, we need to build and update a common ground between the explainer and the explainee: the richer their common ground, the easier it will be to find the correct explanation. In particular, their objective was to estimate the explainee's ToM to make such an estimation more and more similar to the authentic explainee's State-of-Mind (SoM) (Ciatto et al., 2020; Premack and Woodruff, 1978). As they highlighted, an agent's ToM of another agent can be considered an approximation of this latter's SoM (Ciatto et al., 2020). Thus, the more the explainer knows the explainee and the richer their common ground, the easier it will be to provide the proper explanation given the context (Thellman and Ziemke, 2021).

Moreover, Devin and Alami (2016) used robots' ToMs in shared autonomy contexts to reduce unnecessary communication and produce less intrusive robotic behaviours. More importantly for us, they theorised a possible use of robots' ToMs to estimate the lack of information on the robot or understand unexpected users' behaviours. Following their view, the robot's estimation of its collaborator's mental state considers information about the previous and current goals, plans, actions, and perceptual information about the environment. ToM can have different levels, *e.g.*, 2-nd order ToM refers to the ability to recognise that the partner is able themselves to build and use ToMs: those mechanisms have already been observed also in HRI (Matarese et al., 2022).

Similarly, Chakraborti et al. (2017) proposed an idea based on model reconciliation for a similar problem. They hypothesised that the two agents should try to resolve the discrepancies between their internal models to find the most appropriate explanation. Our framework encourages exploiting the user's models to understand what they know and still need to know. Furthermore, being aware of the discrepancies between the explainer and the explainee's internal models could benefit the personalisation of the explanation process by providing hints about the gap to fill.

The idea is similar to the Model Reconciliation (Chakraborti et al., 2017), where the two agents resolve the discrepancies between their internal models to find the most appropriate explanation. In their setting, to explain *"become a process of identifying and reconciling the relevant differences between the models"* (Chakraborti et al., 2017). Indeed, they highlighted the importance of making the AI system acknowledge the differences between its internal models and those of its human collaborators.

Since those models differ, we have the same issue the other way around: for efficient interaction, it is also crucial that the human is aware of the internal states of the robot. Tabrez et al. (2019) focused on users' task understanding to detect incomplete or incorrect beliefs about the robot's functioning. The authors proposed and tested a reward-based framework for estimating reinforcement learning (RL) task understanding. Their goal was to create a shared mental model between humans and robots by allowing the latter to recognise when the human needs additional information. Such situations may occur because humans misunderstand robots' goals and plans due to the differences in their internal models. Task understanding has also been investigated from a dialogical perspective. In particular, Groß et al. (2023) investigated the role of negation in human-robot explanatory dialogues.

Moreover, despite the importance of sharing information between the collaborators in HRI, choosing what information to provide is also crucial to preventing unnecessary communication. A strategy based on situation awareness has been proposed to address this problem by selecting what information, among multiple options, a robot should provide to humans to address their particular information needs (Sanneman and Shah, 2020). The authors proposed a three-level framework to address also different types of explanations (*e.g.*, *what-*, *why-*, and *what if*-questions).

Finally, Stange and Kopp (2023) recognised three main capabilities for explainable robots: (1) the detection of users' needs for explanations, (2) the identification of the situation-specific nature of the explanation needs, and (3) communication abilities to robustly deliver the explanations. Starting from our framework, Stange et al. (2022) designed and developed a dialogical model for explanations in HRI. With their model, the robot can reply to human users' requests for explanations with explanations referring to its internal state. The authors stressed the iterative nature of their model to manage the explanatory processes as dialogues.

The first contribution of this thesis moves forward in the direction of a personalised XAI in scenarios where social interaction mechanisms can enrich the explainability problem. In this regard, we presented a theoretical framework for user-centred XAI in HRI that leverages theories and findings from social sciences to align robot explainability to humans' expectations and habits. Our framework contributes to the field of personalised XAI by providing a general structure for both the explanation interaction and the explanation generation. Moreover, we gave absolute centrality to explainees in the explanation generation phase by exploiting their interaction history with the robot. This way, the framework's objective becomes to provide socially grounded and personalised explanations.

2.3 Explanations in collaborative decision-making tasks

The impact of different XAI techniques on human users during decision-making tasks has been studied mainly in the human-computer interaction (HCI) field (Gambino and Liu, 2022; Lai et al., 2021). Such works include comparisons between different explanation types. For example, Lim et al. (2009) investigated the effects of why and why not-explanations on users' system understanding; they showed that why-explanations led to better understanding and trust towards the system than why-not-explanations. More recently, the efforts moved towards subtler comparisons: *e.g.*, rule- vs example-based explanations with decision support systems to investigate both regarding system understanding and persuasiveness (van der Waa et al., 2021).

Consequently, user-centredness and personalisation drove the XAI research in HCI. Millecamp et al. (2019) showed that users' characteristics (*e.g.*, the need for cognition) influence users when interacting with explainable recommendation systems. However, in a follow-up study Millecamp et al. (2020) found instead that users' openness affects whether they would like to reuse the explanatory system. Similarly, Conati et al. (2021) proved that providing explanations increases users' trust and perceived usefulness and provided insights on personalising explanations using users' personality traits. Furthermore, Tintarev and Masthoff (2012) evaluated personalised explanations and found that those led to higher user satisfaction than non-personalised ones.

When moving to the human-robot interaction (HRI) context, we can find very few studies on XAI with decision-making tasks (Setchi et al., 2020), especially if we are interested in investigating different XAI strategies or customisation and user-centredness. For example, Kaptein et al. (2017) implemented a belief-desire-intention agent on a Nao robot and investigated what explanation style both children and adults prefer. They showed that adults prefer goal-based explanations, while children do not show particular preferences.

Instead, several approaches have been proposed to explain robot planning. Chakraborti et al. (2017) proposed to afford explainability as a reconciliation model using the Fetch robot. Their approach aims to progressively change the human model to bring it closer to the robot's, making the robot's plan optimal for such changes in the human model. Sukkerd et al. (2018) proposed an explainable planning representation to ease explanation generation and a method to generate contrastive explanations as policy justification. Finally, Devin and Alami (2016) moved towards the direction of a user-aware XAI during shared plan execution.

On the other hand, there are several studies with virtual embodied agents (Anjomshoae et al., 2019). Gong and Zhang (2018) proposed an approach to explaining robot behaviour as intention signalling using natural language sentences. They evaluated their method via an online study with a virtual robot. Wang and Belardinelli (2022) proposed using augmented reality to show XAI feedback and the robot's internal beliefs. They tested their method with a virtual Johnny robot and proved that such an approach could enrich the HRI. Differently, Amir and Amir (2018) developed an algorithm to summarise robots' behaviours by extracting information from the agents' simulations. They evaluated their algorithm with a virtual agent in a game-like scenario.

With our first user study, we aimed to answer our RQ1 moving toward user-centeredness and contributing to the discussion about the influence of XAI in HRI. Our experiment involved a complex human-robot teaming task (*Connect 4*) and investigated how two different explanation strategies influenced people in a collaborative decision-making task with a social robot. The first strategy, called *classical* used precise counterfactual explanations to help the robot justify its game choices. Instead, the *shared experience*-based strategy generated counterfactuals from the games that participants and the robot already played, thus exploiting their past experience. This latter explanatory strategy follows our above mentioned framework since it focuses on the human partner's point of view, in particular on the interaction history between the human and robot. Hence, we investigated whether and how a partner-aware explainable robot (the one using the *shared experience*-based explanations) was more persuasive than a robot focused only on the task.

People's personalities have been identified as essential factors in HRI. However, the research on personality is fragmented and lacks a unified framework (Robert, 2018). Hu et al. (2022) demonstrated that several personality dimensions, such as extroversion, affect physical HRI. When dealing with XAI, users' personality dimensions have been studied for several objectives, such as personalization (Weitz et al., 2022). Kallina (2020) investigated whether XAI users' personality dimensions affect how the system influences them. She found that users' personality traits did not impact their tendency to critically review the XAI outcome as their confidence regarding the system's accuracy. Similarly, Abdulrahman et al. (2022) found no influence of users' personality dimensions on their preferences in explanations given by a virtual advisor.

Robots' influence and persuasiveness have been studied for a long time since researchers observed that several influence mechanisms that occur between humans also occur in HRI (Saunderson and Nejat, 2019). Hashemian et al. (2019) explored two persuasive robot strategies based on social power: persuading the users (1) by giving information, or (2) by providing social rewards. Similarly, Saunderson and Nejat (2022) investigated how a robot's persuasive behaviour influences people's decision-making, comparing two persuasive strategies (emotional vs logical) during a guessing game. Moreover, coming to the XAI field, Nayyar et al. (2020) have taken the study on trust and influence with explainable robots to the extremes by investigating the effect of an emergency (simulated) robot's explanations during an evacuation when such explanations differ from the crowds' movements.

The work presented in Chapter 5 answers to our RQ2 focusing on how people's personality traits affected their interaction modalities with explainable robots during complex and collaborative decision-making tasks. We also investigated the learning during the interaction with an explainable robot, intended as the degree of alignment to the robot's behaviour before receiving its suggestions. Although we have been unable to find previous research investigating some learning effects with explainable robots, several studies focus on XAI teaching and the effects of explanations on users' learning in HCI (Fiok et al., 2022). Such a misalignment occurred because the decision-making context is less investigated in HRI than in HCI contexts when involving XAI.

2.4 Evaluating the quality of explanations

Despite the growing number of works regarding the quality of the XAI, it is unclear what constitutes a good explanation. On the one hand, researchers within the field recognise the urgency of designing metrics for XAI (Nauta et al., 2022). On the other hand, it is difficult to find objective metrics. This is because XAI models' efficacy strongly depends on the application context and users' expertise (Hoffman et al., 2018). This is one of the main reasons the community concentrates on the quality of explanations in recent years. For example, Wang and Yin (2021) highlighted three desiderata for XAI systems: to (1) understand the AI model, (2) recognise the uncertainty underlying an AI prediction, and (3) calibrate their trust in the model. Moreover, they compared features contribution and counterfactual explanations on such desiderata. They found that feature contribution explanations satisfy more desiderata for expert users, while counterfactual explanations do not improve trust.

Since explanations are always directed to a human user, most such works assessed XAI systems' properties through user studies. The more interesting ones are addressed to non-expert users (Janssen et al., 2022). These latter involve tasks that need special skills to be performed; skills that ordinary people generally do not have. For example, Lage et al. (2019) used two application domains: a fictional alien's food preferences task and a real clinical diagnosis task. Wang and Yin (2021) also used two peculiar decision-making tasks: recidivism prediction and forest cover prediction tasks. Further, van der Waa et al. (2021) used a diabetes self-management use-case where naive users and the system had to find the optimal insulin doses for meals. Finally, both Goyal et al. (2019) and Wang and Vasconcelos (2020) used image classification tasks in which participants were asked to recognise two bird species.

Several works regarding the assessment or comparison of XAI methods tend to define their own measure of goodness (Lage et al., 2019; van der Waa et al., 2021). However, a method has been proposed to objectively measure the *degree of explainability* of the information provided by an XAI system (Sovrano and Vitali, 2022). The authors proposed an algorithm to quantitatively measure how many archetypical questions a set of information can answer. They assume that the information under study is "good" and form the basis of the explanation provided by the XAI system. Moreover, Holzinger et al. (2020) proposed the System Causability Scale to measure the explanations' quality based on their causability (Holzinger et al., 2019). Finally, Wang and Yin (2022) proposed a different point of view regarding assessing XAI systems' goodness by comparing several types of explanations in different application contexts with respect to three desiderata: to improve people's understanding of the AI model, help people recognise the model uncertainty, and support people's calibrated trust in the model.

Recent works regarding user-centredness focus on users' trust towards the system and highlight context-awareness and personalisation as main approaches to user-centredness (Williams, 2021). Personalisation in XAI has also been implemented by exploiting the users' personality traits and correlating them with users' preferences or behaviours (Böckle et al., 2021; Martijn et al., 2022). Instead, Bertrand et al. (2022b) reviewed a relevant corpus of literature to understand which human biases researchers reflect in their XAI methods (also without noticing).

Most of the works in the XAI field with user studies regard decision-making (Wang and Yin, 2021) or classification tasks (Goyal et al., 2019). The rationale behind this choice is the promise of better performance when coupling human users with expert AI systems (Wang and Yin, 2022). Such a promise is almost always kept, although a few studies show that team performance decreases when using some form of XAI (Schemmer et al., 2022b). Indeed, some authors highlighted that AI advice is only sometimes beneficial mainly because humans have shown to be unable to ignore incorrect AI advice (Ferreira and Monteiro, 2021; Janssen et al., 2022; Schemmer et al., 2022a).

We recognised the lack of objective and quantitative measures for explanations of quality that would involve end-users. Thus, in our last user study, we focused on assessing the informativeness of explanations during a decision-making task. In particular, our assessment task aimed at measuring how much *new information* the XAI system provides to non-expert users. To do so, we designed a task that require skills that generally people do not have - the management of a nuclear power plant - and let non-expert users learn to perform it by interacting with expert explainable agents (*i.e.*, a virtual agent and a humanoid social robot). Moreover, we investigated the impact of different collaborative agents (a virtual agent and an embodied humanoid robot), and compared it to the results achieved by participants who experienced only autonomous learning. In doing so we answered to our RQ3 and RQ4.

Part II

Partner-Aware Explanations in HRI

Chapter 3

A theoretical framework for XAI in HRI

Peter Pan: [to the lost boys] Let's get ready to show them the white light we're made of, boys.
Captain Hook: [to the pirates] Remember the fires of hell that forged you. Charge.
S. Spielberg, Hook

3.1 Introduction

Artificial intelligence (AI) techniques - especially machine learning (ML) ones - are being increasingly used in several areas of our society (Kotsiantis et al., 2006). The growing use of ML systems underlined the need to make them transparent to human users, which means clarifying the rationale behind their decisions to those who use them. The need for transparency in ML became critical when the models were used in crucial areas, such as insurance or credit application (Gunning and Aha, 2019).

To this need, which is becoming increasingly urgent, the AI community has responded by bringing to the fore the explainable AI (XAI) research field (Vilone and Longo, 2020). Among the enormous number of different ML techniques, those based on neural networks have been considered more opaque than others (Adadi and Berrada, 2018). Thus, most of the efforts in XAI have been directed at making neural networks more explainable (Montavon et al., 2018).

However, what an explanation is and what form it should take is not trivial. The first attempts at XAI used explanations that sounded reasonable to the developers, without any scientific support about the effectiveness of such explanations for non-expert users (Mohseni et al., 2018a). This was so because XAI was meant to be used by expert users to debug ML-based technology or improve existing ML models.

Nevertheless, in recent years, several works have been proposed to address the XAI problem in a new social fashion Arnold et al. (2021); Ehsan et al. (2021); Wallkötter et al. (2021). Miller proposed to bring insights from social sciences to build tailored-to-human explanations (Miller, 2019). In his work, he pointed out that XAI is both a computer science and a human-agent interaction problem. People explain themselves to others every day, and artificial agents should exploit their habits in XAI scenarios.

Furthermore, people attribute human-like traits to such agents, especially when they exhibit human-like behaviour or appearance (Phillips et al., 2018b). Thus, people may expect their explanations to comply with the conceptual framework they are used to (De Graaf and Malle, 2017). Hence, it sounds reasonable to exploit people's cognitive mechanisms in XAI interactions (Westberg et al., 2019).

In our opinion, the social aspect of XAI is even more crucial in the human-robot interaction (HRI) field, where the conversational aspects of XAI are stressed (Grice, 1975; Hilton, 1990). Several works have already proposed replicating with robots the exact cognitive mechanisms people use in everyday interactions. Developing those systems considering humans' needs means making XAI human-centred (Hellström and Bensch, 2018; Westberg et al., 2019).

Most proposed solutions have yet to reach a complete human-centred approach (Miller, 2019). Cawsey (1993) has already proposed the idea of integrating a user model into the explanatory system, similar to Wiener's BLAH system (Weiner, 1980). The key idea is that the more the explainer knows the explainee and the more their common ground is rich, the easier it will be to research the proper explanation given the context (Thellman and Ziemke, 2021).

Even though the number of works following this philosophy is rapidly increasing, a more generalist approach is missing in HRI. Indeed, the most complex XAI solutions are ad-hoc in terms of (1) the AI technique they seek to render explainable, (2) the context in which the system is used, and (3) the users who utilise such systems.

Chakraborti et al. (2017) proposed the Model Reconciliation problem, where the two agents resolve the discrepancies between their internal models to find the most appropriate explanation. The appropriate explanation becomes the one that identifies and reconciles the differences between the internal models of the explainer and explainee. This is a dyadic phenomenon: both the explainer and the explainee should be aware of the differences between their internal models, a problem already addressed in HRI to detect incomplete or incorrect beliefs about the robot functioning (Tabrez et al., 2019).

Sanneman and Shah (2020), on the other hand, addressed the problem of selecting the proper explanation through a situation awareness strategy. Among multiple options, the robot should provide an explanation that addresses the users' information needs. From the robot's side, such a framework requires high levels of interaction, personalisation, and awareness of the human partner.

Hence, we aimed to move toward a personalised XAI in scenarios where social interaction mechanisms can enrich the explainability problem. In particular, this work presents a theoretical framework for XAI in HRI. Unlike other works in the field, our framework models both the explanation's generation and the interaction between the parties. We considered the process of explaining as composed of two sub-processes: the explanations generation and providing. The explanations generation phase uses theories and findings from philosophy, psychology and cognitive sciences to model what people expect from explaining agents. The explanation providing phase, instead, regards communication and is highly platform-dependant. Therefore, the framework aims to be partner-aware, which means it generates personalised explanations (Sokol and Flach, 2020) and contextualised ones concerning the specific partner's needs. In particular, it considers the effects of partners' actions, their information needs, and human-inspired heuristics. We focused on *post-hoc* explanations (those asked after the occurrence of the event to be explained) because they are the most common type of explanations in people's everyday lives and the more interesting one from the social-interactive perspective (Lipton, 2016).

This chapter is organised as follows. In Section 3.2, we present our framework and describe its components. Finally, we discuss the implications and limitations of our proposal in Section 3.3.

3.2 The framework

The awareness of the importance of the social-interactive aspect of XAI leads to interesting consequences. First, the explanation process is an exchange of information between agents (Hilton, 1990; Walton, 2004); thus, at least two agents are involved. In this regard, we will use the notation used by Miller (2019) throughout the thesis. Hence, we will refer to who asks for explanations with the term *explainee* and to whom has been addressed by the request of explanation as the *explainer*. In our case, the human takes the role of the explainee, and the robot takes the role of the explainer.

Figure 3.1 shows a graphical representation of our proposed framework. As one can see from the figure, the framework has a twofold objective. On the one hand, it models the

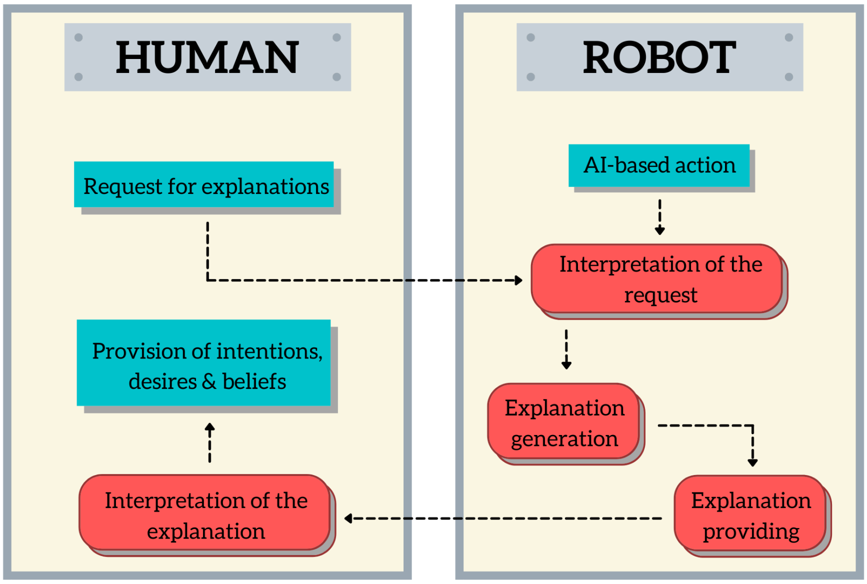

HUMAN
ROBOT
Request for explanations
Provision of intentions, desires & beliefs
Interpretation of the explanation
AI-based action
Interpretation of the request
Explanation generation
Explanation providing

Figure 3.1 A graphical representation of the framework. Since two agents are involved, the human takes the role of explainee, and the other agent, the robot, takes the role of explainer. The blue boxes refer to agents' actions, while the red ones represent the framework's modules. The arrows between the agents indicate communication actions, while the others indicate logical and temporal relations between the modules.

interaction between explainer and explainee as a simple dialectical interplay. On the other hand, the framework shapes the explanations generation process (for details, see Figure 3.2 and Section 3.2.2). The blue boxes refer to agents' actions, while the red ones represent the framework modules that allow both agents to perform the actions. The arrows between the agents indicate communication actions, while those between modules indicate their logical and temporal relations.

As mentioned above, we focused on *post-hoc* explanations (Lipton, 2016). The main characteristic of such explanations is that the explainee asks for explanations as soon as the explainer acts. For this reason, the request for explanations immediately follows the robot's action. Indeed, as we can see from Figure 3.1, the interaction starts from this latter. In particular, unexpected robot actions trigger the human's requests for explanations: people commonly ask for explanations when presented with unexpected or surprising events (Harman, 1965).

Following our framework, the robot, once it receives a request, should interpret it, choose an explanation to give, and communicate it. These three phases - interpretation, explanation generation, and explanation providing - are managed by three distinct modules working in a cascade. The explanation-providing phase completes the transfer of information from the explainer to the explainee. Once the explainer explains, the explainee should translate such an explanation into the agent's intentions, desires and beliefs. The explainee can perform this crucial phase by building a Theory of Mind (ToM) of the artificial agent (Malle, 2006).

3.2.1 The interpreter

Once received a request for explanations, the first action that an explainable agent should perform is to interpret the request. As mentioned above, such requests can be interpreted as why questions (Dennett, 1989; Lewis, 1987; Lipton, 1990). Moreover, they are contrastive questions because something the explainee did not expect has occurred (Hilton, 1996; Lipton, 1990). Conversely, they expected another counterfactual event (Hesslow, 1988; Hilton and Slugoski, 1986; Lombrozo, 2012; McGill and Klein, 1993).

It might happen that, during a complex interaction, the robot does not receive a precise contrastive why question. Hence, during the interpretation phase, the explainee should transform the received question into a contrastive form which we can easily split into two distinct parts: the *fact*, which is the event that occurred, and the *foil*, a not occurred event that the human expected to happen (Lipton, 1990). Due to the contrastive nature of explanations, it is always possible to produce a fact and a foil from a why-question (Hilton, 1990; Lipton, 1990). Thus, the objective of the interpretation phase is to transform the received request for explanations into the question "why *fact* rather than *foil*?".

Generating a fact and a foil could be particularly challenging when the explainee leaves the foil implicit. People are generally good at inferring non-explicit foils; thus, it might happen frequently that the explainee would leave the foil implicit (Hilton and Slugoski, 1986; Lipton, 1990; McGill and Klein, 1993). The interpretation phase should also provide a mechanism to infer such implicit foils to handle these scenarios. Some authors proposed simplifying the problem by defining the foil as the negation of the fact. The resulting interpreted request should be of the form: "why *fact* rather than not-*fact*?" (Lipton, 1990). Other authors claimed we can use the concept of *normality* to address this problem. Thus, since unexpected outcomes elicit requests for explanations, we can see the expected one as the "normal case"; the interpreted request becomes something like "why *fact* rather than the *normal case*?" (Hilton, 1990). Interestingly, this latter solution assumes an interaction history between explainer and explainee; thus, a well-established common ground of experiences and knowledge.

3.2.2 The explanation generation phase

As seen in the previous section, the interpretation phase produces a fact and a foil: these are the inputs of the explanation generation phase. Given an interpreted request for explanations, the goal of this phase is to produce an explanation that would answer the question and sound reasonable to the explainee. In other words, this step aims to find the most appropriate explanation given the fact and foil. For this reason, the explanation generation phase is the framework's core. Figure 3.2 shows a schematic representation of the explanation generation step. It consists of three sub-steps working in cascade: the search for causality, the context setting, and the explanation selection.

The search for causality step aims to find causalities between the fact and foil. It also uses the robot's objectives and the environment's state as a precious source of information. (Borgo et al., 2018; Zelvelder et al., 2021). The first sub-step produces a set of causal chains corresponding to all the possible causal relationships between the abovementioned elements.

The second sub-step aims to find the most appropriate causal chains given the users' specific needs. At this stage, the framework selects the subset of previously produced causalities that best suit the partners' needs. We call such needs the *users' information needs*, meaning the informational needs related to the level of users' expertise in the task (*e.g.*, in collaborative tasks), the level of users' familiarity with the robot, or the interactive history between them; hence, the common ground that they have built during previous interactions.

At the beginning of the third phase, having more than one explanation fitting the context might be possible. In this case, we should select just one to provide a precise and concrete explanation to the explainee. Our framework provides for applying filtering based on human-inspired heuristics to do so.

The search for causality

The first step of the explanation generation phase is understanding what exactly the request for explanations refers to. During this sub-step, the framework searches for causality between the fact and foil to produce a sound explanation. It also uses the robot's objectives and the environment's state to relate the fact and foil. Indeed, the robot's objectives are crucial in its behaviour, particularly its action selection. It is easy to understand why by thinking about tasks in which the robot's actions change the environment (*e.g.*, during RL-based tasks (Matarese et al., 2021b)), or when its previous behaviour determines the robot's possible actions (*e.g.*, in planning tasks). Hence, in pursuing its objectives, the robot might pass through several intermediate steps before landing at the final goal: the human user could not

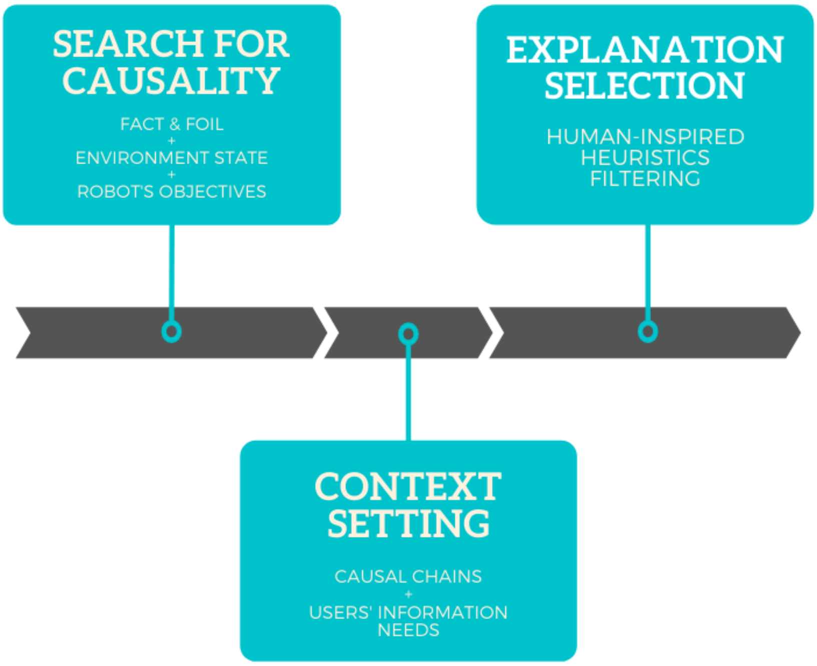


Figure 3.2 The explanation generation pipeline. It is composed of three sub-steps: the search for causality, setting the context, and the final filtering.

find intuitive the usefulness of such intermediate steps. For this reason, taking into account the robot's objectives could help relate the fact and foil in a way that end-users can easily understand.

For similar reasons, the system considers the environment's states in trying to link together the robot's objectives, the fact and the foil. Also this approach can be helpful during planning- or RL-based tasks in different ways: during planning tasks, the robot knows that some states are necessary to arrive at other more interesting ones while, during RL tasks, the relative reward values are a function of the robot's inner states. On the one hand, considering the environment's configurations over time - thus, the changes applied - the system can cover all those scenarios in which the effects of its actions on the environment are informative from an explanation perspective. On the other hand, considering the environment's state means also considering the effect of others' actions in such an environment. A change in the environment's state can also be due to human actions, especially in collaborative contexts. Thus, considering the state of the environment means also considering the effects of human actions.

This sub-module aims to produce causality chains: paths of causes between a set of events (Hilton et al., 2005). Thus, the output of this sub-module consists of a set of causal relationships that could clarify the rationale behind the robot's actions.

The context setting

The second sub-step regards selecting a subset of the causal chains previously produced to find the most appropriate ones. This selection occurs according to the peculiarity of the context that the explainer and explainee share. The chosen explanations should be selected based on what we call *users' information needs*, which refers to the information the partners need according to their level of expertise and their interactive history with the robot.

Hesslow (1988) stated that the criteria for the explanation selection are not arbitrary, bringing as an argument to his thesis the similarities between the explanation selection strategies of different people. However, the peculiarities of the context that the explainer and explainee are facing become particularly relevant in critical situations, such as continual learning tasks (Awasthi and Sarawagi, 2019), failed or uncompleted actions (Leddo et al., 1984), or physical causality (Lombrozo, 2009).

As mentioned in Section 3.1, a request for explanations is always due to a gap between the explainer's actions and the explainee's understanding of the rationale behind such actions. The explainer can determine this gap by knowing or inferring the explainee's information needs. Theoretically, by building a reliable ToM of the explainee, the robot can set the correct explanation goals based on its partner's beliefs and intentions. In its simplest form, such ToM could consist of what the robot already explained to the explainee. However, more generically speaking, the robot might store the information about what the explainee already knows, what part of the environment they are currently perceiving, and anticipate their actions based on their past intentions and current beliefs. This aspect enables the explainer to choose, among all possible explanations, the ones that better fit the explainee's needs.

Moreover, Devin and Alami (2016) used robots' ToMs in shared autonomy to estimate the robot's lack of information or understanding about unexpected users' behaviours. Their objective was to reduce unnecessary communication and produce less intrusive robotic behaviour. Our framework introduces this concept into the explanation process. In our opinion, this is not just customisation but a precise way to put the user at the centre of the explanation interaction. Hence, if the robot has a reliable ToM of the explainee, it could infer what information to provide to fill the gap they have highlighted with their request for explanations.

The explanation selection

If more than one explanation fits the context, final filtering is necessary. This assumption comes from cognitive psychology: explanations are selective (Hilton, 2017; Miller, 2019); thus, people select the explanation that they think perfectly replies to the answer.

We argue that the final filtering should be based on heuristics inspired by cognitive science: *e.g.*, to select recent causes (Miller and Gunasegaram, 1990), or to prefer more controllable causes above non-intentional ones (Girotto et al., 1991). This way, the level of detail is also chosen because, in case of repeated requests for explanations, the explainer can decide to prefer more detailed explanations since the previous one has not been effective enough.

To select what information to provide above multiple choices, a situation awareness-based strategy has been proposed (Sanneman and Shah, 2020): their "Level 1 XAI" includes explanations about what the system is doing and its decisions. Moreover, Sanneman and Shah (2020) stated that, for explainable robots, such a level might include the perceptual information that results in the AI model's prediction. For this reason, we believe a shared perception-based approach could be more suitable for these kinds of scenarios (Kuhl, 1998; Matarese et al., 2022). For instance, shared perception can help during the explanation selection process to choose among several explanation candidates based on the humans' beliefs.

3.2.3 The explanation providing phase

Once the system has a good explanation, all that remains is communicating it. The objective of this module is to choose how to provide the chosen explanation: it defines the explanation's shape and the communication modality. Regarding the latter, it depends on both the robot's communication skills and context: *e.g.*, the robot could provide a verbal explanation or a graphical one via a tablet. Also the shape of the explanation is related to the robot's communication skills: *e.g.*, if the robot can speak, it has to choose the shape of the sentence.

Once the explainer has explained, the explainee needs to interpret it based on the intentionality of the robot's action (ToM). This step is critical because, by putting together their ToMs, they can build a mutual understanding based on common ground, which can improve the robot's explanatory skills. An example of this attempt at alignment can be found in Tabrez et al. (2019), where the authors made the robot explain its reward function to give insights about its actions' rationale concerning the task.

3.3 Discussion

Why should we put so much effort into proposing a theoretical framework for XAI in HRI? In our opinion, the most relevant reason regards anticipating the needs of our society. In recent years, we have seen a spread of AI-based solutions. Companies and individuals increasingly use AI-based systems in the industry, entertainment, and home environments. Consequently, pushed by the introduction of laws and regulations (Adadi and Berrada, 2018), the scientific community developed an increasing interest in XAI themes. Moreover, the need for those AI systems to interact with humans in more complex settings and understand the robots' behaviours is pushing the XAI community to contaminate their work with insights from social sciences.

This work provides a general framework to customise according to the developers' needs and the robots' technical specifications. At the same time, we constrained some framework processes to suit theories and findings from the fields of philosophy, psychology and cognitive sciences. The framework can be used as a theoretical base to develop experiments and XAI models on social robots. Such a development can be done both from the interactive perspective - using the "explanation as a dialogue" template - and from the technical perspective - by implementing the explanation generation pipeline for specific use cases.

The interactive perspective is captured by the loop shown in Figure 3.1. The loop starts from the human's request for explanations, then continues with the robot's interpretation of the request, it goes on with its explanation providing, and ends with the human's interpretation of the explanation. From the robotics point of view, the key steps are the second and third ones. In particular, the interpretation of the request represents an important technical challenge, especially if the robot provides for verbal communication. This means that the interpretation of the human's request should split the verbal request into tokens and recognise facts, foils, and contingent information among those tokens. Although these issues can be addressed with modern large language models, robots' understanding of people's requests still remains an urgent issue to address to provide natural and smooth HRI. On the other hand, the explanation providing strictly depends on the robot's communication abilities and whether it can exploit different forms of communication, such as verbal, non-verbal, and social cues. Several social robots have tablets on them to show additional graphical information and such communication channel can be exploited to, especially when reinforcing the explanations provided with graphical information.

In Chapter 4, we used our theoretical framework and implemented it in a peculiar human-robot collaborative scenario. In particular, we followed its philosophy when considering the human-robot past experiences to generate counterfactual explanations. In that experiment, we stressed the exploiting the common ground that human and robot build during their collaborative interaction. Moreover, in Chapter 7, we counterpoised contrastive explanations to classical causal ones. Contrastive explanations, which provide the exact reasons why an action or behavior occurred *instead of* an expected one, can be seen as the simplest form of implementation of our theoretical framework.

3.3.1 Example scenario

To better understand the peculiarity of our framework, it may be worthwhile to provide an example scenario. Let us imagine a domestic robot helping its owner, Bob, assemble a piece of furniture. Since it is the first time for both in such a task, they follow the instructions in the user manual to organise their work and develop a plan of action. While the person reads the instructions, the robot stores such information and builds a plan through AI models.

In other words, we have a complex collaborative task: a building task resolvable with planning methods. The agents' actions are constrained by rules, which are the assembly instructions that regulate, among other things, the order of actions to be performed. We can say that both agents need to communicate, build a common ground of knowledge (*i.e.*, the instructions), and be aware of each other's actions to perform efficiently such a task.

At a certain point, Bob focuses on his work when the robot passes him the box with the 5cm-long screws. Bob is confused because he does not need such screws now nor in the next step in which, according to the instructions, he will need the 10cm-long ones. Thus, he asks the robot "why did you pass me the 5cm-long screws?". The robot has to interpret the request for an explanation since Bob did not provide the foil but only the fact. Since it has no information other than the assembling instructions, it interpreted the request as "why did you pass me the 5cm-long screws rather than *one of the other* screws?". Then, during the context setting phase, the robot selects the explanations related to the 10cm-long screws because it realises that Bob will need those in the very next step of the task. Consequently, it replies "because I can not see the 10cm-long screws". This way, Bob understands that by passing the tools needed in the next step, the robot is thinking ahead and, more importantly, aligned with his work. He looks up and immediately sees the 10cm-long screws behind a piece of furniture; thus, he can infer they are occluded to the robot's view.

3.3.2 Limitations

In this section, we highlight the limitations of our framework by discussing typologies of systems for which our approach is not suitable. In particular, we focus on systems that need *legibility* rather than explainability, and on those XAI systems with a specific use and do not allow complex interactions with their human users.

Broad attention has been given to robots' legibility and the difference between this latter and explainability (Lichtenthäler and Kirsch, 2016). To summarise, robots' legibility refers to the ability to be not ambiguous to human observers. In contrast, robots' explainability refers to the ability to produce explanations for their behaviour (Chakraborti et al., 2019). This substantial distinction brings practical differences between systems we would expect legibility and those we ask for explanations.

For example, one would demand legible robotic behaviours exactly when such behaviours take place. On the contrary, explanations are asked after unexpected robotic behaviour to understand its rationale or before such behaviour to understand the functioning of the robot's model. In this regard, systems that need legibility rather than explainability can not use our framework mainly because legibility answers questions different from explainability. Moreover, the simplicity that often characterises the scenarios in which robots that need legibility are used constitutes another reason why our framework is unsuited for such robots.

For example, we want robotic arms in factories collaborating with human workers. They have to perform specific tasks; thus, we can not expect the robots to build an interaction history with human workers (unless we want to make the robot capable of providing personalised assistance). Indeed, legibility is often related to robot motion (Dragan et al., 2013), while explainability is more often related to other tasks, such as learning. In such scenarios, we need legibility more than explainability; thus, the goal of an agent is transparent and predictable in its movements, meaning that we can easily infer its intentions from them.

To discuss the other typology of systems for which the proposed approach is inappropriate, let us take diagnosis systems as a case study. Because of the growing use of ML models in every engineering field, those systems are becoming increasingly integrated with AI solutions and, in recent years, XAI models to help users during their maintenance tasks. Diagnosis systems involve scenarios in which expert users ask the system simple queries to check if everything works as it should, detect biases in datasets, or check the latter's integrity. Hence, the limitations of this context make it impossible to pursue a complex interaction between users and XAI systems. They do not need a common ground to exploit to understand the system better because the "communication" is de facto one-way and has one objective: to query the system and inspect its answers to detect any malfunctions. This problem applies to all those scenarios that provide rigid interaction protocols between users and XAI systems. Since the core of our framework exploits the interaction exchange between explainer and explainee, it can not suit these kinds of XAI systems.

Part III

Partner-Aware Explanations During Decision-Making Tasks in HRI

Chapter 4

The role of explainable robots' persuasiveness

"Perhaps she would respond to an alternative form of persuasion."
G. Lucas, Star Wars: Episode IV - A New Hope

4.1 Introduction

Since the '70s, artificial intelligence (AI) techniques have been increasingly used in our society. In particular, during the last two decades, machine learning (ML) systems have become growingly present in our daily lives, often without us noticing (Shinde and Shah, 2018). Such a growth impact on people's lives requires an effort to make these systems explainable to non-expert users to increase trust and acceptance (Vilone and Longo, 2020), and encourage a conscious approach to the use of such technology (Nesset et al., 2021; Sanders et al., 2014b).

Parallel to the growth of ML, great strides have been made in robotics, and a strong interest in human-robot interaction (HRI) issues has developed (Sheridan, 2016). When ML and robotics merge - *e.g.* when ML manages robots' behaviour or their interaction with humans - the need for explainable AI (XAI) becomes even more crucial because people often attribute human-like traits to robots (Phillips et al., 2018a), even the more complex characteristics, such as the 2nd-order theory of mind (Matarese et al., 2022).

The HRI context is particularly suitable for a social and user-centred XAI because of two main reasons (Miller, 2019; Tabrez and Hayes, 2019). On the one hand, it has been shown that people quickly adapt their interaction habits to robots (Ahmad et al., 2017). On the other hand, we expect that robots will have long-term and personalised interactions with us (Belgiovine et al., 2022). In other words, the HRI field is pushing towards robust customisation of robots' behaviour; in our opinion, this translates into social and user-centred XAI (Matarese et al., 2021a).

Although the number of studies investigating user-centred approaches to XAI has increased in the last years, we can find very few articles about the effects of such approaches in social HRI contexts (Anjomshoae et al., 2019; Liao and Varshney, 2021). Moreover, almost all of them are placed in the human-computer interaction field (Abdul et al., 2018). In particular, they compare different XAI techniques with human-AI teams performing decision-making tasks. They also investigate the persuasiveness of such XAI techniques and their impact on the perception of the underlying AI system (Paleja et al., 2021a).

The impact of different XAI techniques has been studied mainly in human-computer interaction (HCI) (Gambino and Liu, 2022) for decision-making tasks (Lai et al., 2021). For example, Wang and Yin (2021) compared features contribution and counterfactual explanations with respect to three desiderata, while van der Waa et al. (2021) compared rule- and example-based explanations with a decision support system.

In the literature, we can also find several approaches towards personalised XAI. User characteristics, such as the need for cognition (Millecamp et al., 2019) and personality traits (Millecamp et al., 2020), have been used to provide user ad-hoc explanations. The increase in trust justifies such a growing interest in personalisation, perceived usefulness (Conati et al., 2021), and user satisfaction (Tintarev and Masthoff, 2012) with personalised XAI systems.

When moving to the human-robot interaction (HRI) context, we can find very few studies on XAI Setchi et al. (2020) and users' preferences (Kaptein et al., 2017). For example, explainable robot planning has been approached with a reconciliation model (Chakraborti et al., 2017) or using contrastive explanations as justification for the robot actions (Sukkerd et al., 2018).

In this study, we proposed comparing two XAI approaches in a social HRI scenario: we called them *classical* and *shared experience*-based. We set the HRI as a social decision-making task in which participants and the robot had to collaborate to beat the computer at the *Connect 4* game (Figure 4.1). We used counterfactual explanations because of the complexity of the ML model needed to solve the game and their contrastive nature (Malle, 2006). The *shared experience*-based counterfactuals comprised of game configurations that the robot retrieved from the games that it and user have already experienced; thus, they reflected a partner-aware approach to explainability. In this sense, such an explanation strategies implemented the philosophy of the theoretical framework we presented in the previous chapter. Contrary, the classical approaches produced more precise counterfactuals.

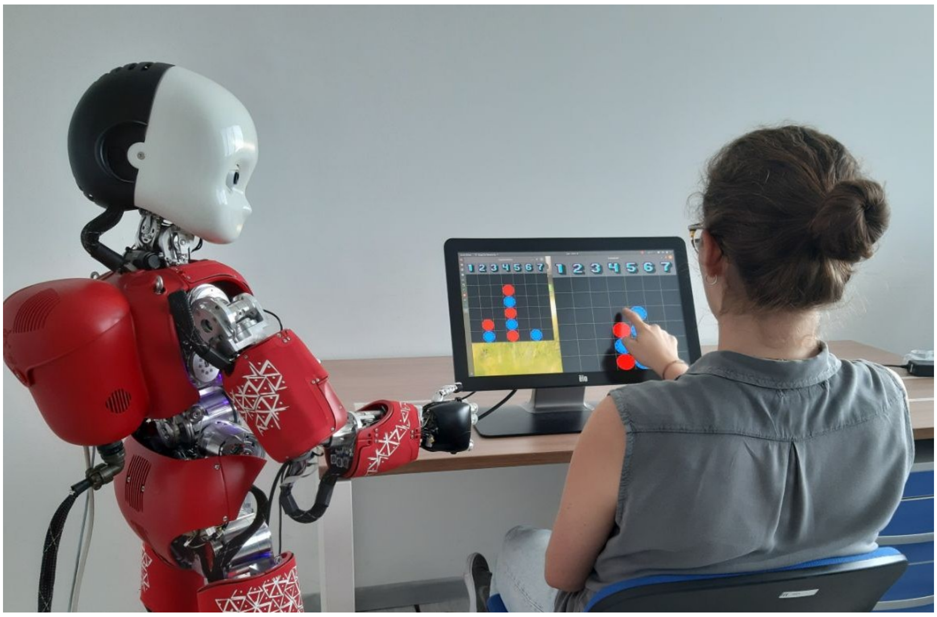


Figure 4.1 The experimental setup of the collaborative Connect 4 game. The bigger window on the screen represents the board game, while the smaller one contains the *counterfactual* of the robot's last explanation. In this case, iCub is showing the counterfactual for the fact *"play in column 4"* and foil *"play in column 6"* as the robot's opening strategy was to build the structure shown in its explanation. As a result, the participants followed iCub's suggestion.

Our aim was to test whether *shared experience*-based counterfactuals were more effective than *classical* ones regarding performance, persuasive power and perception of both the robot and self.

4.2 Methods

This study aimed to examine how different counterfactual generation approaches affect people's task performance, perception of the robot, and persuasive power in a social HRI. We designed the interaction as a competitive game where the human-robot team plays against the computer at the *Connect 4* game.

We were also interested in studying a non-explanatory robot; thus, participants faced three phases corresponding to the experimental conditions (Figure 4.2).

- *Solo*: participants played alone against the computer.
- *No exp*: participants and iCub played together, but the robot produced no explanations.

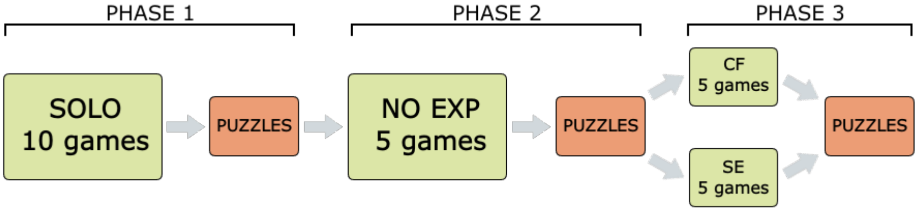


Figure 4.2 The experimental design schema. All participants (22 total) performed the *solo* and *no exp* phases; then, they were split into two groups (11 per group): one faced the *CF* explanation phase, while the other the *SE* explanation phase.

- *Exp*: participants and iCub played together, and iCub produced explanations. With half of the participants, iCub produced classical explanations (*CF* group), for the other half *shared experience*-based explanations (*SE* group).

As a result, we obtained a within-subject experimental design where each groups of participants addressed all the different experimental phases.

4.2.1 Procedure

After reaching the experimental room with iCub, we illustrated to participants the experimental protocol (the three phases), and instructed them to play *Connect 4* with iCub (as a team) against the computer. Then, we briefly recap the rules of the game[1], and we instructed them on how to use the touch-screen and application (Figure 4.1). Moreover, we told them what form the robot's explanations had and how to interpret them: we did not mention anything about the experimental groups not the difference between the two explanation strategies. Regarding iCub, we told them it knew the game rules and that we were testing a new algorithm in a learning phase. The experiment had three phases; the human-robot team was always the first to play.

During the *solo* phase, iCub could not intervene in the games; it could only watch the games and comment on their results: it would say "Oh no, we lost!" in case of a losing game, or "Yeah, we won!" in case of a victory. iCub turned its head to look participants in the eyes each time it talked to them, then went back to the screen to keep following the games. Participants had to play ten games against the computer during such a phase. Right after the matches, they had to solve 20 puzzles. The puzzles were configurations of the *Connect 4* game in which participants had to play only one move, which they thought was the best. The

[1] https://en.wikipedia.org/wiki/Connect_Four

puzzles presented after the *solo* condition were predetermined and with increasing difficulty. We used these puzzle as a punctual baseline of the participants' performance.

During the *no exp* phase, iCub could participate in the decision-making process of the five games. The decision-making process was similar to the one used in (Wang and Yin, 2021). Participants first indicated their choice, and then iCub told them whether it agreed. If not, it told the participants which column it would drop the fiche instead. At the end of this interaction, participants made the final decision and, finally, moved. Also in this phase, participants had to solve 20 puzzles. However, rather than being predetermined as in the *solo* condition, the application took them from the matches of the current phase. In particular, the application randomly chose the puzzles among the configurations in which participants opted for a move different from iCub's. We used these puzzles to investigate whether and how participants could replicate iCub's in-game suggestions rather than reproducing their own moves.

The third phase (*CF* or *SE*) was similar to the second one but with explanations. iCub produced an explanation each time its choice differed from that of participants. Such explanations were displayed by showing the counterfactual on another smaller window next to the board game (Figure 4.1). iCub accompanied the explanations with an arm movement, indicating such a new window. As before, participants had to solve 20 puzzles (taken from the current phase's games) and fill out a questionnaire. We used these puzzles as we did for the second phase.

4.2.2 AI models

We equipped both iCub and the computer with ML models. From the computer's side, we chose a Monte Carlo Tree Search (MC-TS) algorithm (Browne et al., 2012). Using such a model, we could perform satisfactorily and adjust its depth to accommodate the difficulties of playing against it. We set the search depth we used during the experiments with a preliminary pilot study in which we asked our colleagues (10 total) to play alone against the computer. We chose the MC-TS model's depth and time limit (sim. number = 10, max iter. = 2000, timeout = 2), which allowed our colleagues to win $\simeq 50\%$ of the time.

On the other hand, we made iCub play via a deep neural network (DNN) to obtain semi-optimal performance against the computer. We chose the AlphaZero architecture (Silver et al., 2018) trained to play the *Connect 4* game because it could easily reach perfect performance against the MC-TS agent. We fine-tuned the model to reach 90% of victories against the computer's MC-TS model because we wanted a semi-perfect iCub.

4.2.3 XAI model: counterfactual generation

The complexity of the game required a complex model. Typically, the more complex the AI model, the less transparent and explainable (Barredo Arrieta et al., 2020; Došilović et al., 2018). We chose counterfactual explanations to obtain a good tradeoff between explainability and performance since we could not renounce Alpha-Zero's DNN complexity. In particular, we used example-based counterfactuals: when iCub disagreed with participants' first choice, it showed a board configuration - the counterfactual - to justify its move. The chosen counterfactual had two peculiar properties: (1) it was similar to the current configuration of the game, but (2) it was different enough to induce iCub to take the participant's initial choice.

Although there are several methods to produce counterfactual explanations from a DNN (*e.g.*, DiCe from Lundberg and Lee (2017)), we preferred to build them manually because of the impossibility of post-hoc filtering. We needed it because we needed counterfactuals showing legal configurations; in other words, configurations of the game that could result from a legal match. Indeed, current counterfactual generation techniques assume that all the input features are independent. This is not the case for *Connect 4* since the value of a board's slot strictly depends on the value of the slots directly under it.

Thus, we decided to use counterfactuals that would satisfy our requirements. By letting the DNN play against the computer, we collected a dataset of more than 11000 configurations (without duplicates) with the move the DNN would take in each of them. Consequently, we organised these configurations depending on such moves.

```
Algorithm 1 Counterfactual generation
Require: $1 \le fact, foil \le 7$; $fact \neq foil$
  $counterFacts \leftarrow counterFactsDataset[foil]$
  sort $counterFacts$ w.r.t. $fact$ using $L_1$ norm
  $counterFact \leftarrow counterFacts.top()$
  return $counterFact$
```

Algorithm 1 illustrates how we retrieved the counterfactuals from the dataset, where the iCub's move is the fact and the user's is the foil. We used the L_1 norm as a measure of similarity (Lundberg and Lee, 2017; Mothilal et al., 2020), which ensures good properties for our needs (Wachter et al., 2017). Through this ordering, we ensured that the retrieved counterfactual was the configuration most similar to the current one among those in which iCub would make the user move.

On the one hand, we used the dataset mentioned above to implement the *classic* approach. On the other hand, we collected the counterfactual dataset from the matches participants played during the experiment to implement the *shared experience*-based one. Those matches comprehended also the first ten games (the solo phase) in which iCub could not intervene. In particular, we saved into the dataset the configurations in which participants moved differently from what iCub would. Those datasets had an average size of $\mu = 198.1$ ($\sigma = 18.5$) configurations. This led to two main differences between the counterfactual types:

- The mathematical precision (L_1 norm) was higher for *classical* counterfactuals than for *shared experience*-based one because of the large difference between the two datasets' sizes.
- There was a high probability that the participants had not previously encountered the former, whereas they had already encountered all of the latter.

Classic and *shared experience*-based counterfactual shared the same shape and meaning, they were both legal configuration of the game representing in which circumstances iCub would do the participants' actions rather than its'. Hence, no one could notice any difference without knowing they were from different datasets.

4.2.4 Participants

We recruited 22 participants, all signed informed consent before the experiment, approved by the ethical committee of "Regione Liguria". 19 participants were 20-30 years old, with the remaining participants ranging from 30-45; 14 identified as women, 7 as men, and 1 preferred not to specify.

4.2.5 Measures

We split the experimental measures into three macro-groups: performance, persuasive power, and perception of the self and robot. In this chapter, we focused on the first two, while results regarding participants' perception of the self and robot are presented in the following chapter.

Performance

We measured participants' performance during the games of all the three phases in terms of the percentage of winning games, and during the puzzles after the *solo* condition in terms of the percentage of correct moves.

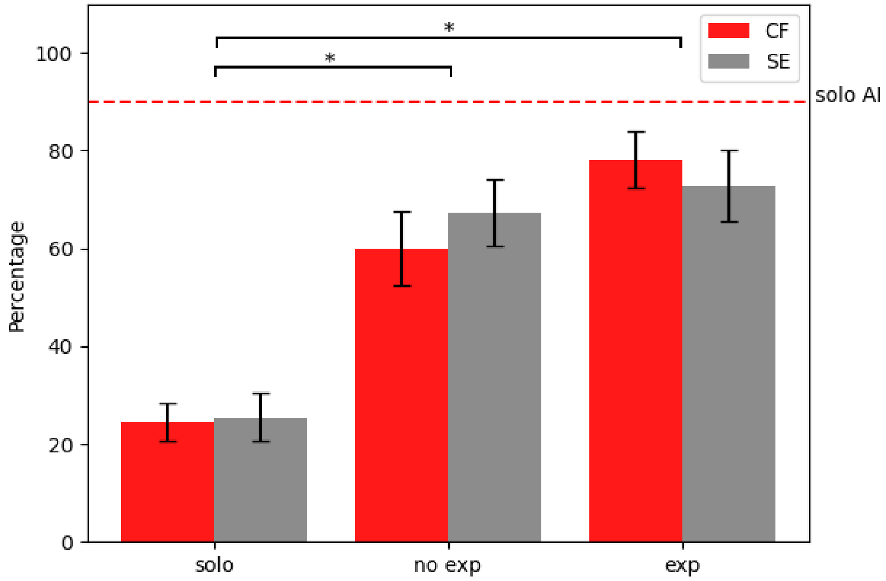


Figure 4.3 Mean and standard error of participants' performance of the two groups (CF vs SE) in each condition. The red dashed line refers to the robot's AI performance. We can see that the human-robot team outperformed humans alone; however, they did not reach the AI performance.

Persuasive power

We measured whether participants confirmed their initial choice or changed in favour of iCub's one during phases 2 and 3. Moreover, through the subsequent puzzle phases, we measured how much participants unconsciously "absorbed" iCub's suggestions, *i.e.*, implicitly learned from previous iCub's suggestions and then replicated its choices during the puzzles.

4.3 Results

4.3.1 Performance

We measured the game performance as the percentage of games won against the computer (Section 4.2.2). Figure 4.3 summarises the average task performances for each condition between the two groups (CF vs SE). With a two-way mixed-model ANOVA analysis, we determined that there was a significant difference between the performances of the groups ($F(4,104) = 39.72$, $p < .001$). A post-hoc test with Bonferroni correction showed that solo

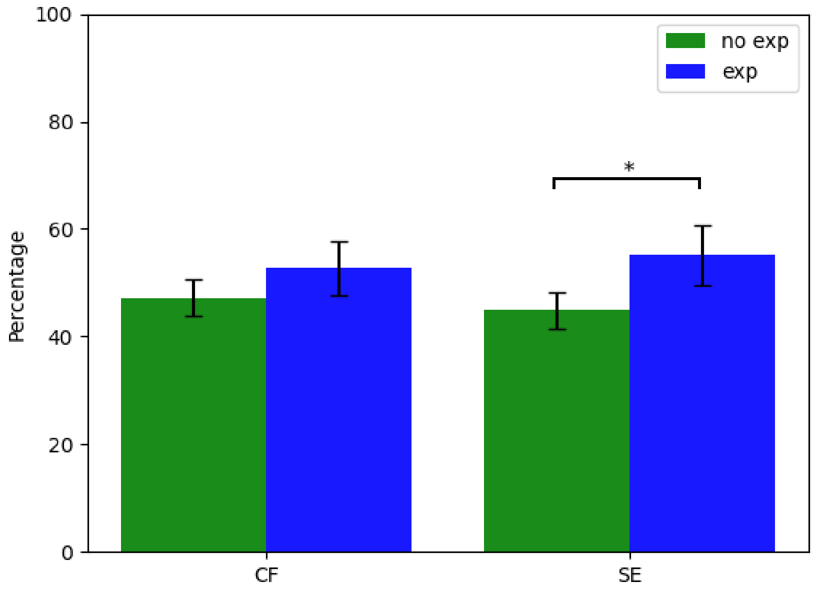


Figure 4.4 Mean and standard error of the percentages of *equal* moves of the two groups (CF vs SE) divided by experimental condition (with vs without explanations). Here, we consider only the *equal* moves so we can see that SE participants replicated iCub's moves during the explanation phase significantly more than during the previous one. Contrary, we found no effects within the CF group in this regard.

AI performance significantly differed from all the groups in all conditions ($p < .001$ for all such comparisons). Similarly, the *solo* (human) performance was statistically different from those of all the groups in all conditions ($p < .001$ for all such comparisons).

The number of puzzles correctly solved after the solo condition was comparable between the two groups (CF vs SE). Indeed, participants in the CF group solved $\mu = 14.8$ puzzles ($\sigma_{\bar{x}} = .85$), and those in the SE group solved $\mu = 14.27$ puzzles ($\sigma_{\bar{x}} = .59$).

4.3.2 Persuasive power

To measure iCub's persuasive power, we collected three types of moves:

- *Equal*: participants chose since the beginning the move that iCub would have chosen (*e.g.*, participants chose and made a move 2 when also iCub would have chosen it).
- *Follow self*: participants insisted on playing their move instead of iCub's one, *e.g.*, participants chose move 2, iCub chose the 3, and participants made move 2.

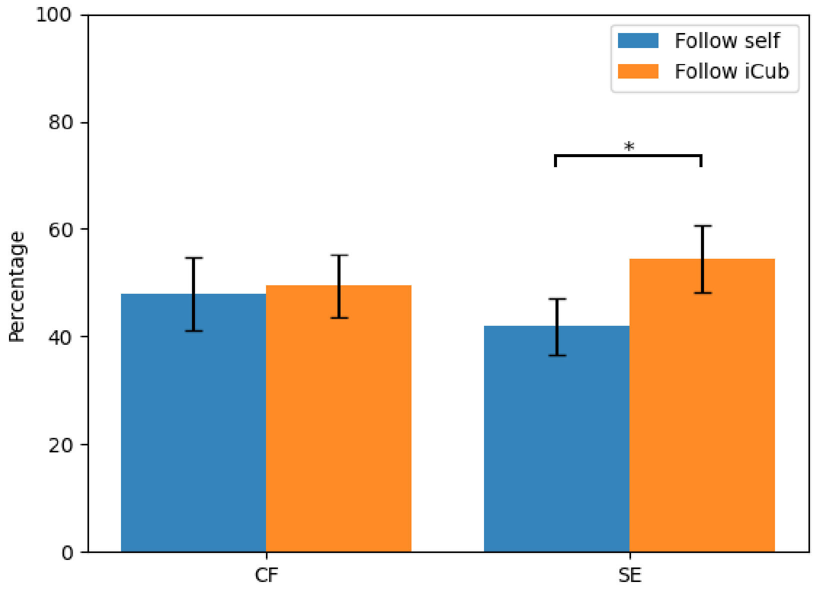


Figure 4.5 Mean and standard error of the percentages of move types (follow self vs follow iCub) of the two groups (CF vs SE). Here we do not consider other types of move. We can see that SE explanations significantly improved the robot's persuasiveness, while CF explanations did not.

- *Follow iCub*: participants opted for iCub's suggestion, *e.g.*, participants chose move 2, iCub chose the 3, and participants made move 3.

All the other move types' percentages were negligible. We averaged the percentages of such move types over the games because we found no particular patterns depending on the games' numbers.

Figure 4.4 shows the average percentage of participants' *equal* moves between the conditions with and without explanations. A two-way mixed-model ANOVA test revealed a significant difference between the conditions ($F(2,217) = 7.06$ $p = .001$). A post-hoc test with Bonferroni correction showed a statistical difference between the conditions *no exp* and *exp* only for the SE group ($p = .03$).

Figure 4.5 shows the average percentage of participants' move types during the *exp* condition for both groups (CF vs SE). A two-way mixed-model ANOVA test showed a significant interaction between the conditions and types of moves ($F(2,324) = 3.84$, $p = .022$). A post-hoc test with Bonferroni correction showed that the percentage of *follow self* moves is significantly lower than the *follow iCub* ones ($p = .009$) for the SE group.

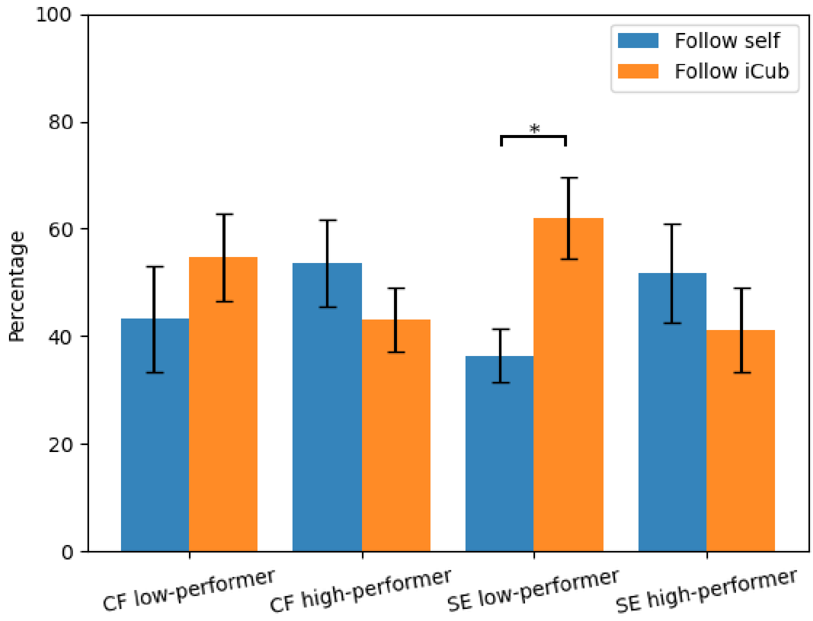


Figure 4.6 Mean and standard error of the percentages of the move types of participants divided by performance in *solo* condition. Low performers won less than the group average during the *solo* condition, while high performers won more than the average of the group of participants. Here, we do now consider other types of moves.

Figure 4.6 shows the average percentages of participants' move types divided by performance during the *solo* condition. We considered *low-performers* participants who won in *solo* condition less than the average of the entire group; the remaining were considered as *high-performers*. In particular, participants were divided into: *CF low-performers* (6 total), *CF high-performers* (5 total), *SE low-performers* (7 total) and *SE high-performers* (4 total).

A two-way mixed-model ANOVA test revealed that there was a significant interaction between the conditions and types of moves ($F(3,212) = 6.02$, $p < .001$). A post-hoc test with Bonferroni correction showed that only for the SE group, between low-performers, the percentage of *follow self* moves was significantly less than the *follow iCub* ones ($p < .001$).

Figure 4.7 shows the average number of move types made during the puzzle phases divided by participants' tendency to follow iCub suggestions during the main game. Participants in the low-follower category followed iCub less than 50% of the time; otherwise, we considered them high-followers. The resulting division was: *CF low-followers* (5 total), *CF high-followers* (6 total), *SE low-followers* (5 total) and *SE high-followers* (6 total).

A two-way mixed-model ANOVA test showed an interaction between the conditions and types of moves ($F(3,36) = 21.49$, $p = .013$). A post-hoc test with Bonferroni correction showed that only for the SE group, low-followers reproduced their moves more than followed iCub's in-game suggestions ($p = .013$).

4.3.3 Perception of the robot and self

We discarded 2 participants from the analyses of the questionnaires because they failed to reply to attention checks designed to measure respondents' engagement quickly. 90% of them declared to have already participated in a study with iCub, and all the items we submitted to the participants reached a good Cronbach's Alpha ($> .65$), indicating that the scales were all acceptably reliable.

We found no differences between the groups in the items regarding the robot's warmth, competence and human-likeness. Similarly, participants' level of satisfaction with the explanations did not present differences between the groups. A repeated measures mixed-model ANOVA test showed that the answers to the IOS test were significantly higher in the post-experiment questionnaire than in the pre-experiment one ($F(1,19) = 6.72$, $p = 0.018$). Furthermore, participants' perceived level of goodness in playing the game grew through the phases ($F(2,36) = 12.06$, $p < .001$). In this regard, a Bonferroni post-hoc correction showed a significant difference between phases 2 and 3 ($p = .007$) and 1 and 3 ($p < .001$). Likewise, participants' perceived difficulty of the game grew through the experimental phases ($F(2,36) = 9.39$, $p < .001$). A post-hoc test with Bonferroni correction showed significance only between phases 1 and 3 ($p = .007$).

4.4 Discussion

Both game and puzzle performance in phase 1 showed that the two groups of participants were homogeneously skilled, allowing a fair comparison between the two groups (Section 4.3.1). Participants' performance is similar to that reported in the human-computer interaction literature regarding human-AI teams in decision-making tasks. Specifically, participants alone performed worse than AI (25% vs 90%) and, with the robot, they did not reach the AI performance, as in (Bansal et al., 2021). We speculate that the reason for this effect has to be found in how the game develops, in addition to the difficulty for participants to identify iCub's failures (Poursabzi-Sangdeh et al., 2021). It happened that iCub gave incorrect suggestions to participants because of its sub-optimal performance. However, participants did not notice those errors since they were subtle and not so evident. Indeed, the more serious errors regarded the first moves of the game. The first moves are the more important ones, and

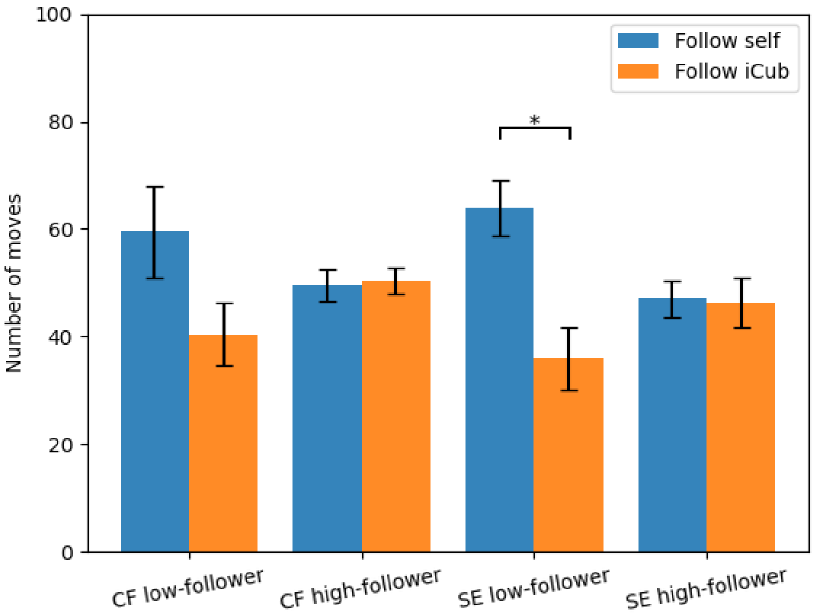


Figure 4.7 Mean and standard error of the number of move types of the participants during the puzzle phases, divided by their tendency to follow iCub's suggestion during the main game. All puzzle phases included 20 puzzles. Low-followers followed less than 50% of iCub's in-game suggestions, while high-followers followed more than 50% of those. Here we do not consider other types of move. We can see that low-followers were not able to replicate iCub's in-game suggestions during the puzzles, while high-follower ones stood at 50-50%.

failing some of those could seriously compromise the entire game. Moreover, participants performed worse without explanations than with them (independently of the explanation type), but the difference in performance is not statistically significant, as also shown by Bansal et al. (2021). Finally, the two counterfactual generation strategies brought similar performance.

Differently from other studies (van der Waa et al., 2021), we found no significant differences in the robot's persuasiveness between the condition without explanations and the two explainable ones (Section 4.3.2). Participants tended to confirm themselves more (and follow iCub less) with classical explanations (CF group) than with shared experience-based ones (SE group). However, participants in the SE group followed iCub more than confirmed their moves, whereas we did not find this difference in the CF group. Thus, shared experience-based explanations brought a higher persuasiveness than classical ones. This could mean that presenting explanations from a common ground gives higher reliability to the robot, making it easier for people to follow it.

Interestingly, this effect was even more relevant for those who performed less than the group average without the robot's help. The robot's higher persuasiveness with low-performer participants gives us insights into the potential danger of letting non-expert users interact with expert explainable robots (van der Waa et al., 2021). Several potential dangers have already been pointed out, such as the difficulty for people to detect AI errors (Poursabzi-Sangdeh et al., 2021), and the biases that guide their interaction with AI systems (Green and Chen, 2019a). We can easily transpose these to our social robots; hence, we still need a profound discussion about the ethical implications of using such technologies in everyday life (Green and Chen, 2019b).

In both conditions, as they played with iCub, participants adapted their way of playing to the robot's. Indeed, let us look at how the percentage of *equal* moves changes through the experimental phase. We can see that it grows between the condition without explanation and the two consequent explanatory ones. However, such a difference is significant only for the SE group. It might be that the reason for such results is in the higher persuasiveness outlined before.

Puzzle phases represented a precise way to measure participants' implicit learning of iCub's strategies. In the SE group, those who were low-followers during the main game tended to confirm their moves more during the puzzles than when solving puzzles the way iCub would. This tendency, although not significant, is also true for the CF group. This result suggests that it was difficult for low-followers to "absorb" iCub's choices because they did not follow them; as a result, they could not replicate such moves during the puzzles.

Participants rated their closeness to the robot higher after the experiment than before, regardless of the group (Section 4.3.3). Thus, participants perceived a social interaction and enjoyed it, and if we also consider how they perceived iCub as part of their group ($\mu = 4.68 \pm \sigma = 1.42$ on 7 points Likert scale), we can say that they considered iCub as a teammate, although they were focused on the task.

The main limitation of this study rely on the simplicity of the interaction modalities and on the fact that the robot provided suggestions and explanations to every move. Indeed, it gave participants no freedom to choose when to receive suggestions and explanations and when not to. On request suggestions and explanation have been investigated in the study we present in Chapters 7 and 8.

In this study, we have seen how participants' over-reliance on the robot's suggestions shaped their behavior and performance of the human-robot team. Although participants performed better with iCub than alone, their team performance did not reach the AI's accuracy. This blind reliance on AI-generated suggestions highlights issues related to people's agency during human-AI collaboration. Moreover, participants receiving the shared experience-based explanations tended to follow the robot's suggestions more than those who received the classical ones. However, these former did not realise that the counterfactuals that iCub showed them resulted from their common ground; thus, we speculate that they unconsciously acknowledged that and this resulted in a higher reliance. In the following chapters, we investigate deeper the role of participants' personality traits in shaping their behavior with the explainable robot, and ask ourselves whether people prefer such partner-aware approaches to explainability because of their inherent usefulness.

Chapter 5

The role of explainees' personality traits

"Maybe that's what a person's personality is: the difference between the inside and the outside."

J. S. Foer, Extremely Loud and Incredibly Close

5.1 Introduction

Explainable AI (XAI) will play a crucial role in AI-powered personal or domestic robots (Vilone and Longo, 2020). Those robots are meant to have long-term interactions with their users (Kaptein et al., 2017). Moreover, people tend to attribute to (explainable) robots complex human traits (Hellström and Bensch, 2018), such as high-level theory of mind mechanisms (Matarese et al., 2022). Hence, the research effort in this domain should focus on personalized and user-aware XAI in human-robot interaction (HRI) to respond to people's high expectations towards the robots' capabilities. Indeed, we can exploit the mechanisms whereby people project human-like traits into robots to ensure smooth interactions that adapt to the users (Matarese et al., 2021a). The long-term goal will be obtaining human-robot reciprocal awareness (Sciutti et al., 2018), also with explainable robots (Tabrez and Hayes, 2021).

One way to implement user personalization exploits their personality dimensions (Conati et al., 2021) and previous experiences (Matarese et al., 2023). In the human-computer interaction (HCI) field, similar customized approaches have already been proposed (Völkel et al., 2019) - also using XAI technologies (Putnam and Conati, 2019) - during human-AI decision-making tasks (Paleja et al., 2021b). Furthermore, users' peculiar characteristics have also been studied in HRI context (Esterwood and Robert, 2021; Robert et al., 2020); however, they have yet to be approached with explainable robots.

Robots' influence and persuasiveness have been extensively investigated, observing that several influence mechanisms between humans also occur in HRI (Saunderson and Nejat, 2019). Robots' persuasive behaviour has been studied in classical decision-making (Saunderson and Nejat, 2022), but also with respect to robots' explainability (Nayyar et al., 2020). Moreover, personality traits have been identified as crucial factors for effective HRI (Robert, 2018). They have been investigated to study how they affect physical HRI (Hu et al., 2022) or for personalisation purposes (Weitz et al., 2022).

Our work addressed the human-AI teaming context with a socially explainable robot to investigate whether users' personality dimensions affect the HRI during such complex tasks. This study compares participants' behaviour during a decision-making task involving an explainable and non-explainable robot. We set the experiment as a social human-robot teaming task where participants and the iCub humanoid robot have to collaborate to beat the computer at the *Connect 4* game (Figure 4.1). We tested how participants' personalities and previous experience with iCub and the game shape the HRI with an explainable and non-explainable robot during decision-making. Finally, we studied how participants tended to align with the robot's way of playing the more the interaction continued.

5.2 Methods

This study builds from the one presented in the previous chapter, focusing the dichotomy of "explanatory vs non-explanatory" robots during collaborative decision-making tasks. We investigated on how users' personality dimensions and previous experiences could shape the HRI. The experiment was composed of three phases corresponding to the experimental conditions (Figure 4.2):

- *Solo*: participants played alone against the computer.
- *No exp*: participants and the non-explanatory robot played together.
- *Exp*: participants and the explanatory robot played together; with half of the participants iCub produced *classical* explanations (CF), for the other half *shared experience*-based explanations (SE).

5.2.1 Participants

All participants ($n = 22$) signed the informed consent approved by the Regione Liguria ethical committee. Nineteen participants were 20-30 years old, with the remaining participants ranging from 30-45; 14 identified as women, seven as men, and one preferred not to specify.

5.2.2 Measures

We can split our experimental hypothesis into two macro-categories: the first (**H1**) regards users' personality dimensions and previous experiences shaping the HRI, while the second (**H2**) regards users' learning. In particular, we formulated a total of three hypotheses:

- (**H1.A**) users' personality and (**H1.B**) their previous experiences affect the HRI with explainable robots; we expect to replicate influence mechanisms that characterise the human-human interaction also in our HRI setting, *e.g.*, an active role of participants' extroversion and agreeableness in accepting iCub's suggestions.

- (**H2**) the more participants interact with the robot, the more they can learn to act like it.

We considered three experimental measures: perception of the self, perception of the robot, and learning; all the items we used are available in the appendix. We asked participants to answer all items on a 7-point Likert Scale.

To investigate participants' behaviour during the decision-making process, we collected three types of moves:

- *Equal*: participants chose the move that also iCub would have chosen before the robot's suggestion.

- *Follow self*: participants confirmed their initial choice even though iCub suggested another one.

- *Follow iCub*: participants opted for the robot's suggestion.

Perception of the self

Before the experiment, we submitted to participants a brief measure of the Big Five Personality Domains (Gosling et al., 2003): Extraversion, Agreeableness, Conscientiousness, Emotional Stability and Openness to Experience. We then submitted to them the Sense of Agency scale, which measures consciously perceived control over one's mind, body, and immediate environment: we divided it into Sense of Positive and Negative Agency scales (Tapal et al., 2017). Hence, we considered participants' personality dimentions those resulting from the Big Five test and the Sense of Agency one. Moreover, we asked whether they had seen the iCub robot or interacted with it. Between the experimental phases, we asked participants to rate how much they thought they were good at playing the game. We also asked them to rate the game's difficulty.

Perception of the robot

In the pre-experiment questionnaire, to evaluate the robot, we also showed the participants an institutional video[1] of iCub and we asked them to indicate their level of accordance with several items regarding their impression of the robot. They evaluated the robot by answering the Warmth and Competence (Fiske et al., 2007) (i.e., how much they felt the robot was competent and warm), Agency and Experience (Gray et al., 2007) (i.e., how much they thought the robot could act and feel) and Anthropomorphism scales (Ferrari et al., 2016) (i.e., how much they thought the robot is similar to a human being). They also answered a short version of the Likeability Scale (adapted from (Spaccatini et al., 2019)), (i.e., how much they liked iCub), and the Perceived Enjoyment scale (how much they enjoyed interacting with iCub) (Heerink et al., 2010b). We then asked them to answer three items adapted from Team Identity Scale (Heere and James, 2007) to evaluate how much they felt about being in the same team with iCub. Finally, we submitted the Inclusion of Other into the Self (IOS) test (Aron et al., 1992). After each game, we asked participants who they thought was more responsible for the game's outcomes (*Who do you think contributed most to the outcome of the last play: iCub, you or both?*). After the third game session, we submitted to the participants six items on their satisfaction with the explanations (Conati et al., 2021). In this last phase, we also submitted to them the Anxiety scale (i.e., how much they were scared of interacting with the robot during the experiment) (Heerink et al., 2010b). After the experiment, we asked participants to answer the same questionnaire to evaluate the robot administered before.

Learning

We considered the percentage of *equal* moves as a direct measure of participants' learning. Hence, by measuring the percentage of such types of moves, we measured how much participants aligned with the iCub's way of playing, meaning they could reproduce the

[1] https://www.youtube.com/watch?v=3N1oCMwtz8w

	Cronbach's α	
Scales	**Pre**	**Post**
Sense of Positive Agency	.80	-
Sense of Negative Agency	.75	-
Warmth	.65	.85
Competence	.78	.77
Agency	.79	.67
Previous Experience	.94	.93
Likeability	.74	.82
Anthropomorphism	.66	.67
Explanations	-	.79
Anxiety	-	.64
Perceived Enjoyment	-	.80
Team Identity	-	.68

Table 5.1 Cronbach's α values for questionnaire's scales.

robot's moves without needing suggestions. Since iCub played significantly better than participants, we considered this alignment an improvement for them.

5.3 Results

All the results that follow, except those referring to Cronbach's αs (Table 5.1), refer to the *exp* phase, regardless of the type of explanations iCub produced. Indeed, the two types of explanations we used produced no differences; hence, from now on, we consider the *exp* phase a unique block.

We performed a preliminary analysis to check our scales' and subscales' internal consistency and reliability (Table 5.1). As we can see from the table, all scales and sub-scales have a sufficient Cronbach's α (Cronbach, 1951). We considered the Warmth (Pre), Agency (Post), Anthropomorphism, Anxiety, and Team Identity scales as reliable even if they have lower α values due to the exploratory goal of the study.

We followed the principle of parsimony by Tabachnik and Fidell (2007) to reduce variables. We considered participants' behavioural measures as dependent variables: the percentage of*follow self*,*follow iCub* and *equal* moves. On the other hand, we considered

Covariates	Variables (%)	Sig.	η^2	Obs. Power
Negative agency	Follow iCub	.021	.346	.678
Agreeableness	Follow self	<.001	.643	.994
Agreeableness	Follow iCub	<.001	.750	1
Exper. w/ iCub	Follow self	.005	.474	.885
Exper. w/ iCub	Follow iCub	.007	.441	.841
Diff. alone	Equal	.002	.518	.932
Diff. alone	Follow self	.002	.520	.933

Table 5.2 Values from the MANCOVA test. This table should be read with the subsequent one to understand the direction of the effects. We can see that participants' negative agency positively affected their tendency to follow iCub; also their agreeableness positively affected such tendency, and negatively their tendency to rely on their own choices.

participants' personality dimensions and their previous experience with the robot and the game as covariates.

Thus, we performed a MANCOVA test with all our dependent variables and covariates; then, we progressively discarded from the analysis those covariates that did not significantly impact such variables. In the last analysis, all the remaining covariates affected at least one dependent variable (Table 5.2). Finally, we checked the sign of the correlation to understand whether the covariates positively or negatively affected the dependent variables (Table 5.3). As a result, we obtained that participants' sense of negative agency, agreeableness, previous experiences with the robot, and the perceived difficulty of playing alone significantly impacted their behaviour during the decision-making task with the iCub robot. Table 5.2 summarises those covariate and dependent variable pairs; moreover, Table 5.3 shows the signs of such effects (positive or negative).

Regarding hypothesis H1.A, according to which people's personality traits impact the HRI with an explainable robot, we found that participants' negative agency had a positive effect on their percentage of *follow iCub* moves as covariate ($F(1,20) = 6.86$, $p = .02$). Instead, participants' agreeableness had a positive effect on their percentage of *follow iCub* moves ($F(1,20) = 38.92$, $p < .001$) and a negative effect on their percentage of *follow self* moves ($F(1,20) = 23.42$, $P < .001$) as a covariate. As an example, Figure 5.1 shows the correlation between participants' agreeableness and their percentages of *follow iCub* and *follow self* moves. We did not obtain a full negative correlation between those two types of moves because participants also performed *equal* moves.

	Move types		
Variables	**Follow iCub**	**Follow self**	**Equal**
Neg. agency	.105	-.05	-.076
Agreeableness	.62*	-.385	-.072
Exper. w/ iCub	.054	-.081	.101
Diff. alone	-.041	-.177	.363

** p <0.01*

Table 5.3 Pearson's ρ values of the correlations between the covariates and dependent variables reported in Table 5.2.

Regarding our hypothesis H1.B, according to which previous experiences with iCub impact the HRI with the explainable robot, we found that participants' involvement in previous experiments with the robot iCub negatively affected their percentage of *follow self* moves ($F(1,20) = 11.71$, $p = .005$) and positively affected the percentage of *follow iCub* moves ($F(1,20) = P10.25$, $p = .007$). On the other hand, participants' perceived difficulty in playing alone against the computer negatively affected the percentage of both *equal* ($F(1,20) = 13.96$, $p = .002$) and*follow self* moves ($F(1,20) = 14.06$, $p = .002$).

As indicated at the beginning of this section, the results presented above refer to the *exp* phase. Our final result in this respect is that none of the findings regarding the effect of the participants' personality dimensions on their behaviour during the task occurred during the *no exp* phase. Furthermore, we found no effect of the participants' previous experience with iCub on their playing style.

Regarding our hypothesis H2, according to which participants would learn to play like iCub, we performed an ANCOVA analysis: we determined that there was a significant difference in the percentages of *equal* moves in the different experimental phases ($F(2,45) = 21.49$, $p < .001$) (Figure 5.2). A post-hoc test with Bonferroni correction showed that such percentages significantly differed in all the phases ($p = .012$ between *solo* and *no exp*; $p < .001$ between *solo* and *exp*; $p = .003$ between *no exp* and *exp*). Hence, the more participants played with it, the more they learned to reproduce iCub's strategy (before receiving his suggestions).

An ANCOVA analysis showed that there was a significant difference in the percentage of victories through the experimental phases ($F(2,56) = 12.35$, $p < .001$). A post-hoc test with Bonferroni correction showed that such percentages significantly differed between the *solo*

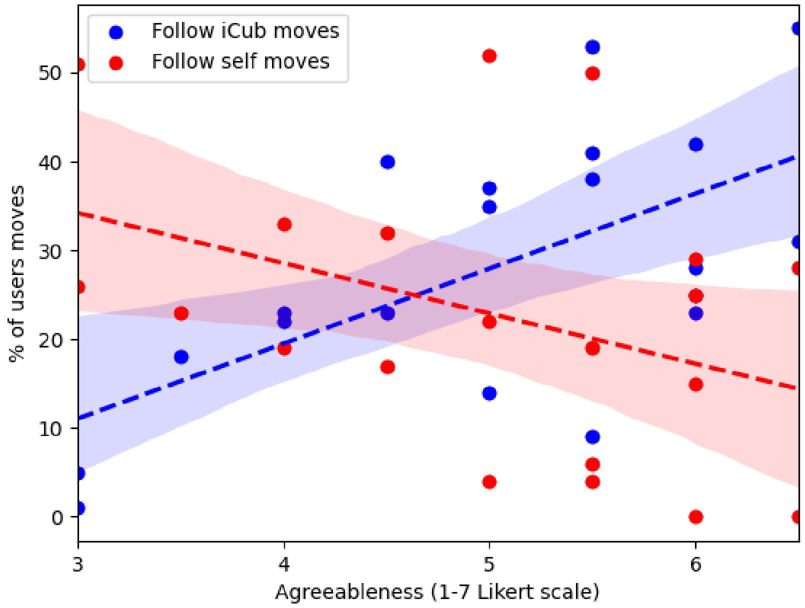


Figure 5.1 Correlations between the averages of participants' *agreeableness* and their *follow self* and *follow iCub* moves during the *exp* phase. Participants' agreeableness positively correlated with their tendency to follow the robot's suggestions, and negatively with their tendency to rely on their own choices.

phase and the other two ($p < .001$ for both). In this regard, the percentage of *equal* moves had an effect as covariate ($F(1,56) = 6.23$, $p = .016$).

Finally, only during the *solo* phase, we found a correlation between the percentage of victories and equal moves (Pearson's $\rho = .603$, $p = .005$). Moreover, the percentage of victories during the *no exp* and *exp* phases positively correlate with the percentage of *follow iCub* moves during such phases (Pearson's $\rho = .498$, $p = .025$ and Pearson's $\rho = .454$, $p = .044$, respectively). On the other hand, the percentage of victories during the *no exp* and *exp* phases negatively correlate with the percentage of *follow self* moves during those phases (Pearson's $\rho = -.594$, $p = .006$ and Pearson's $\rho = .482$, $p = .031$, respectively). Thus, while the participants' victories were due to their skills when playing alone, the human-robot team's chances of winning were related to the persuasiveness of iCub rather than the participants' learning or skills.

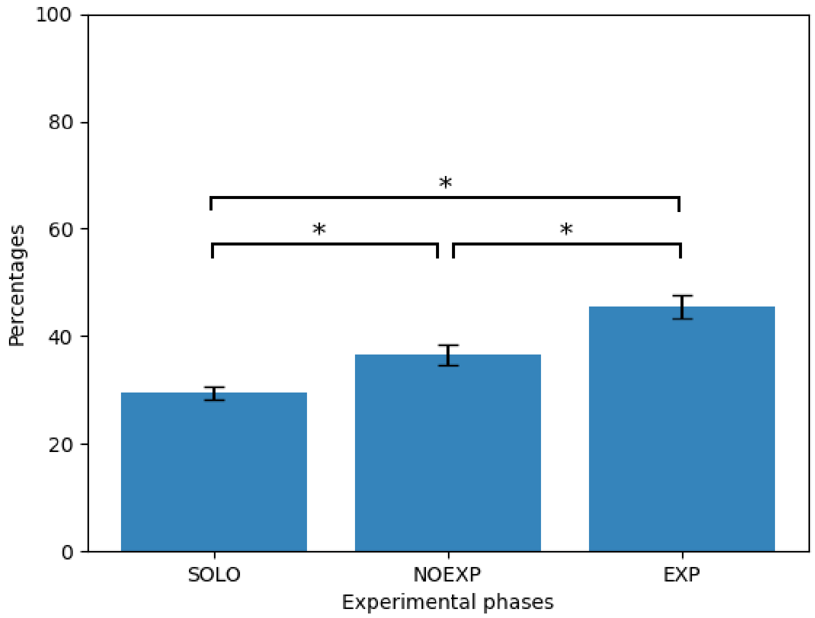


Figure 5.2 Average and standard error of the percentage of *equal* moves through the experimental phases.

5.4 Discussion

This study explored to which extent participants' personality dimensions influenced the interaction with an explainable robot. Moreover, we explored how interacting with an explainable expert robot affected participants' learning. In the original study, we considered two types of explanations (Chapter 4). However, in the analyses presented in this chapter, we found no distinction between those two and then we considered the C-XAI and A-XAI groups as one. We speculate that this effect was due to the subtle difference between the two explanation strategies we exploited: a difference big enough to induce different participants' behaviors but not enough to be affected by their personality dimensions.

A reduction in people's sense of agency has already been observed during HRI (Ciardo et al., 2018). Similarly, our participants' negative agency negatively affected their independence during the task with the explainable iCub. Indeed, we found that such a personality dimension positively correlated with the percentage of *follow iCub* moves. Thus, the higher their negative agency, the higher their tendency to rely on the robot's justified suggestions.

Moreover, agreeableness was the personality dimension that impacted participants' behaviour the most in our decision-making task. As already discussed in the literature, agreeableness has been associated with higher trust towards security robots (Lyons et al., 2020) and closeness (Pantecouteau and Passera, 2017). Further, Gessl et al. (2019) found that agreeableness is the personality dimension with the highest number of associations with technology acceptance.

As we could expect from interaction dynamics within humans, our participants' agreeableness positively impacted their tendency to follow iCub's justified suggestions. Consequently, it reduced their tendency to rely on their initial opinions. It has not to be excluded that the role of agreeableness in accepting iCub's suggestions rather than the own moves has a foundation in reciprocity or social acceptance mechanisms (Zonca et al., 2021). Another reason can be that participants' agreeableness enhanced their trust toward the explainable robot (Sanders et al., 2014b). Our results suggest that we can use personality dimensions to determine the impact of the explainability of the robot on people.

However, it is crucial to specify that our results are based on a brief measure of personality domains, which thus has diminished psychometric proprieties (Gosling et al., 2003). Moreover, personality dimensions and sense of agency indexes were obtained through self-report scales. Further investigation should focus on in-depth the role of personality dimension, not only relying on self-report scales.

Having already performed experiments with the robot iCub impacted our participants' tendency to accept its suggestions. Indeed, having had previous experience with it positively correlated with the percentage of *follow iCub* moves, and a negative one with the percentage of *follow self* moves. Similarly, Gessl et al. (2019) found that experience with technology was among the dimensions that had the most significant relationships with technology acceptance. Our results also confirm the impact of previous experiences in perceiving explainable robots more reliably during decision-making tasks. In our opinion, as Haring et al. (2016) already discussed, previous interactions with the robot iCub influenced participants' perceptions of it.

Also participants' perceived difficulty in playing alone against the computer impacted their tendency to accept the robot's suggestions, as it negatively correlated with the percentage of *follow self* moves. The less complicated participants perceived the game, the more they tended to rely on their initial choices.

Our experiment highlighted an increasing learning effect through the experimental phases. Such learning effect played a crucial role in participants' performance since the percentage of *equal* moves significantly acted as a covariate in the percentage of victories between the *solo* and the other two phases. This "playing as iCub" effect helped participants beat the computer in the *solo* phase: their performance in this phase significantly correlates with the percentage of *equal* moves.

Instead, their performance during the second and third phases positively correlated with the percentage of *follow iCub* moves. This suggests that robot persuasiveness played a significant role in the other two phases.

Before concluding, we need to discuss the limitations of our study regarding the order of the experimental conditions. Since we did not counterbalance the experimental conditions, we can impute this learning effect not only to the robot's explainability but also to the time that participants spent playing with iCub. Furthermore, counterbalancing the order of the explanatory and not-explanatory phases may highlight the role of the experimental conditions' ordering. Hence, removing the ordering effect can strengthen our results. Finally, we need to highlight the limitations coming from using self-reported questionnaires and a brief measure of participants' personality traits like the one proposed in Gosling et al. (2003).

To summarise, interacting with an expert robot enables a learning effect: users tend to align with the robot more and more throughout the interaction. However, this learning effect did not affect their performance in the phases in which the robot provided suggestions. When playing with the robot, they probably understood that their chances to win were linked to how much they followed the robot's suggestions rather than to their acquired skills. Moreover, unlike HCI works, we found no effects of participants' personality in modulating their learning (Conati et al., 2021; Ghai et al., 2021). Participants aligned to iCub's way of playing the game, regardless of the personality dimensions they reported in our questionnaire.

This chapter aimed at highlighting the potential issues that link people's personality traits and their behavior with explainable expert robots during collaborative decision making. We found that the more delicate personality dimensions, such as negative agency and agreeableness, proved to significantly impact participants' tendency of relying on the robot's suggestions. We also observed a learning tendency: people mimicked the explainable iCub's playing style more and more through the experimental phases. In the following chapters, we focus on explanations' goodness rather than on their persuasiveness. In particular, we investigated whether partner-aware explanations are objectively better than classical ones. To do so, we used a novel objective assessment task to measure the explanations' informativeness for non-expert users.

Part IV

Partner-Aware Explanations During Decision-Making Learning-by-Doing Tasks

Chapter 6

A learning-by-doing decision-making task to measure the information power of explanations

"As long as you don't choose, everything remains possible."
J. Van Dormael, Mr. Nobody

6.1 Introduction

The number and complexity of user-centred approaches to XAI have been increasing in recent years (Williams, 2021). The application contexts of such approaches are various. They go from computer applications aiming at providing personalised teaching (Cohausz, 2022; Embarak, 2022) to human-robot interaction (HRI) contexts in which the robot maintains users' models to provide explanations tailored to them (Matarese et al., 2021a; Stange et al., 2022).

These works show a clear trend: user-centred XAI positively affects the interaction between users and systems. It brings to higher users' willingness to reuse the system (*e.g.*, with recommendation systems (Conati et al., 2021)), robots' persuasiveness during human-robot decision-making tasks and human-AI teams performance (Schemmer et al., 2022b).

One of the reasons why the XAI research field is moving toward user-centred approaches is to improve the *goodness* of the explanations. Alongside producing personalised XAI, recent works also aim to evaluate the goodness of the explanation produced. What is meant for “explanation goodness” is still vague; nonetheless, there is a broad consensus about the dependency of this concept on the application context. So far, researchers have measured the goodness of an XAI system used in human-AI decision-making tasks by relying on indirect measures, such as team performance, systems’ persuasiveness, or users’ ability to predict the AI’s decisions.

The lack of objective and quantitative measures contributes to the difficulty of rigorously comparing two XAI strategies regardless of their particular application contexts. Since all the XAI systems have been examined with respect to their application scenario, it is also hard to generalise the results of such research (Gilpin et al., 2018). Indeed, the performance of explanations is rarely tested, and most test rely on heuristic measures rather than explicitly valuing explanations from a human perspective (Ghassemi et al., 2021).

Several works assessed the properties of XAI systems through user studies. They are addressed to non-expert users (Janssen et al., 2022), with both invented and highly specialised domains (Goyal et al., 2019; Wang and Vasconcelos, 2020). Decision-making tasks are the more investigated by those works (Wang and Yin, 2021); unfortunately, works regarding the assessment or comparison of XAI methods tend to define their own measure of goodness (Lage et al., 2019; van der Waa et al., 2021).

Nonetheless, Sovrano and Vitali (2022) proposed a method to objectively measure the *degree of explainability* of an XAI system. Moreover, Holzinger et al. (2020) proposed the *System Causability Scale* to measure the quality of the explanations based on their notion of causability (Holzinger et al., 2019). Differently from the previous studies, Wang and Yin (2022) proposed to address the explanations’ goodness by comparing different types of XAI with respect to three desiderata: model understanding, model uncertainty recognition and calibrated trust.

However, the main aim of an XAI system is to provide the user with information about the functioning of the underlying AI model. So far, the amount of information an XAI system provides has been assessed through indirect measures, such as users’ ability to simulate the AI behaviour. Indeed, recent surveys (Mohseni et al., 2018b) and systematic reviews (Nauta et al., 2022) highlight the need for more objective and quantitative measures to assess the goodness of XAI techniques.

In our opinion, it is worth taking a step back and trying to measure the goodness of XAI systems from the information they can provide to users. Although measuring how much information explanations can generate is challenging, we can focus on the amount of *new knowledge* about the task that such flow of information creates in the users’ mental models. Hence, if we let only non-expert users interact with XAI systems, we can assume that the knowledge they acquired arose from interacting with the system. Finally, if such knowledge is quantifiable, we can assess how much the system has been informative.

We need to introduce an objective and quantitative assessment task to measure a critical factor in XAI: explanations' *information power*. With this term, we mean the amount of information that an XAI system provides about: (1) the underlying AI models' (general) functioning, (2) the reasons behind a particular model's choice, or (3) what the system would do in other circumstances. Under the assumption that the goodness of an XAI system reflects the accuracy of the users' mental models about the underlying AI system (Hoffman et al., 2018), we want to understand whether user-centred explanations are more informative than those that do not take the explainee in consideration.

In this study, we proposed an assessment task to objectively and quantitatively measure the goodness of XAI systems during human-AI decision-making tasks, intended to determine how informative they are to non-expert users. Moreover, we used such a task to study the effectiveness of contrastive partner-aware XAI in both HCI and HRI scenarios. In particular, we were interested in understanding whether a partner-aware explanation strategy is more informative than a classical one.

It has been theorised that not only the explainer is responsible for the explanation generation. Since the explanatory process is social and dialogical (Miller, 2019), the explainee has several responsibilities regarding the effectiveness of the explainer's explanations. Furthermore, Rohlfing et al. (2020) stated that explainer and explainee *co-construct* the explanations, highlighting the active role of explainees.

Following their model (Figure 6.1), we aim to define the explainer *monitoring* of the explainee as the necessary condition for a partner-aware XAI. Hence, our experimental hypothesis regards the partner's level of involvement in the explanation generation process. In particular, the more users are involved in the explanation generation process:

- **H1:** the more information power the XAI system has (resulting in more accurate users' mental models about the task).
- **H2**: the more they become independent of the artificial agent's suggestions and explanations.
- **H3**: the higher is the users' willingness to reuse the XAI system.
- **H4**: the more positive the users' feelings toward the artificial agent are.

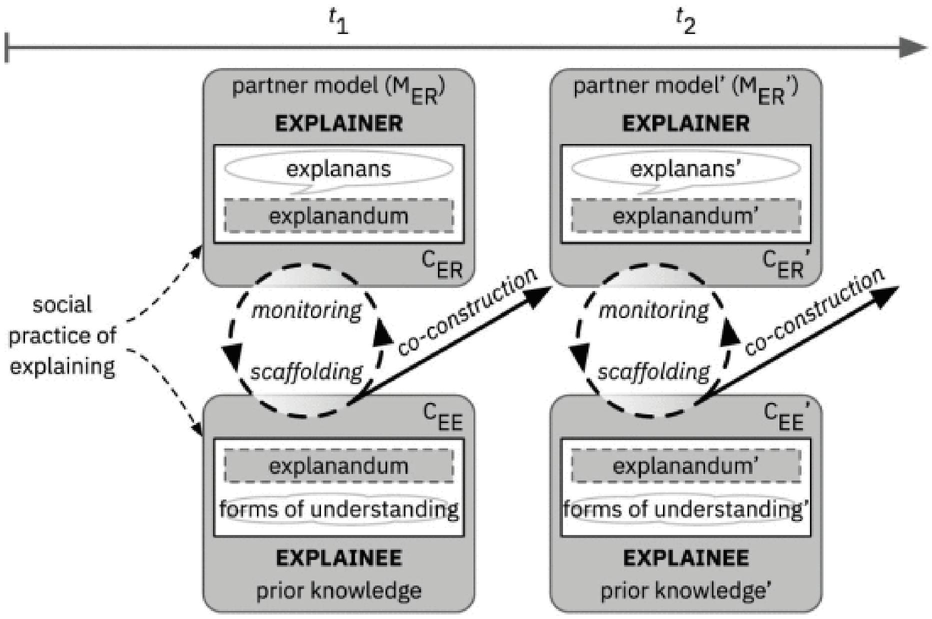


Figure 6.1 The co-constructing approach to the explaining process (Rohlfing et al., 2020). The actors' explanation behaviours adapt by monitoring and scaffolding each other. In our setting, the ER monitors the EE through feedback about their actions and questions.

6.2 Methods

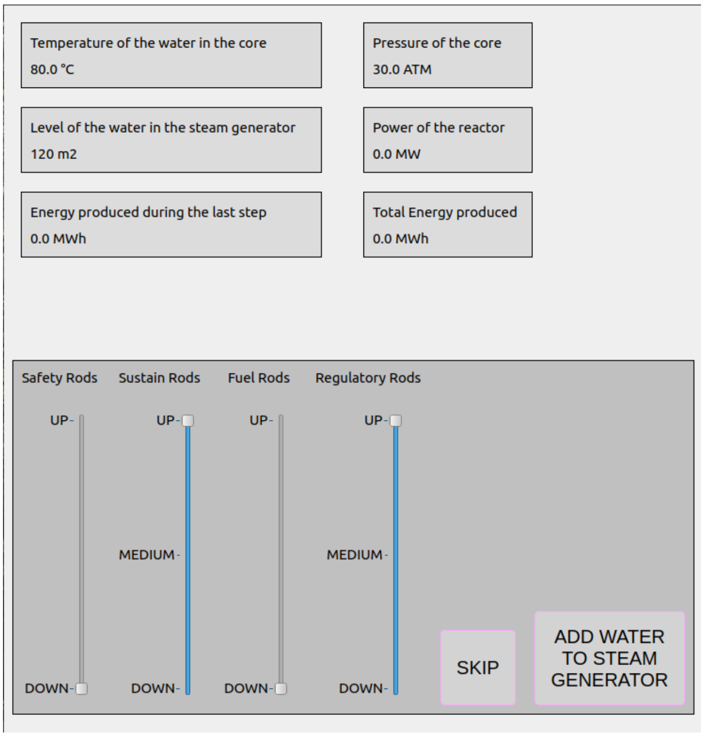
We want to introduce an objective and quantitative assessment task to investigate the information power of two different explanation strategies. Hence, we plan to measure the XAI systems' information power directly and validate the task through a user study.

6.2.1 The task

The assessment consists of a decision-making task where users can interact with a control panel (Figure 6.2) to perform actions in a simulated environment (see Section 6.2.2). During the task, users can interact with an expert explainable AI agent by asking *what* it would do in each step. Moreover, they could also ask *why* it would perform a specific action. Users start the task without knowledge, besides the instructions about interacting with the control panel and the agent.

In a fixed amount of time (30 minutes), the users have to discover:

- Which is the task at hand (*e.g.*, what are the goals of the task).
- The rules that govern the simulated environment.
- The rules that the AI model uses to select its actions.



Temperature of the water in the core
80.0 °C

Pressure of the core
30.0 ATM

Level of the water in the steam generator
120 m2

Power of the reactor
0.0 MW

Energy produced during the last step
0.0 MWh

Total Energy produced
0.0 MWh

Figure 6.2 The nuclear power plant (NPP) application's control panel that participants used during the experiments.

To do so, they could interact with the environment by performing actions and checking the results. Moreover, they could interact with the artificial agent to obtain additional information from its suggestions and explanations.

Subsequently, participants have to complete an *assessment* phase. During such a phase, participants had to perform the same task they learned during the training, but instead of learning how the task works, they had to perform it at their best in a fixed amount of time (10 minutes).

At the end of the task, users have to undergo a test about the task's objectives and rules to assess their level of understanding. The experimenter can choose any type of test (*e.g.*, open or multiple-choice questions), and it should aim to assess:

- Users' level of knowledge about the task's objectives.
- Users' level of knowledge about the task's internal rules.
- Users' ability to generalise the skills acquired during the task.

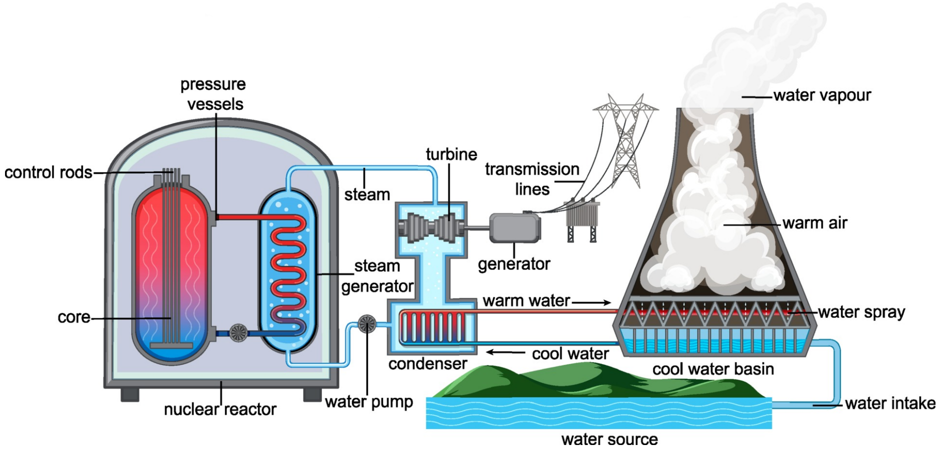


Figure 6.3 The schema of a pressurised water reactor that we implemented in our simulated environment. This schema considers only the control rods, but we also allow the users to manage the other three types of rods: the fuel, sustain and regulatory ones.

Interaction modalities

The roles of the human-robot team and their interaction modalities are simple. The robot can not perform actions, but its role is limited to assisting users during decision-making. However, the robot can not take the initiative to give suggestions either, but it always replies to explicit users' questions. Thus, only the users can interact with the control panel and act in the simulated environment.

Characteristics of the task

We needed non-expert users to consider the information passed through the interaction as new information. Thus, we considered only participants without knowledge about the task and its underlying rules. For this reason, we implemented a nuclear power plant (NPP) management task (Figure 6.3). We chose this kind of task because it met all the requirements we needed: it is challenging and captivating for non-expert users, simple rules govern it, an AI model can learn those rules and, usually, people know nothing about the functioning of nuclear power plants.

The main objectives of the task (which we hide from users) are to generate as much energy as possible and maintain the system in an equilibrium state. The features of the environment are subject to rules and constraints, which we can summarise as follows:

- Each action corresponds to an effect on the environment: thus, a change of its features' value.

- Several preconditions must be satisfied to start and continue nuclear fission and produce energy.
- Some conditions irremediably damage the NPP, bringing to *anomalies*.

6.2.2 The simulated environment

Features of the environment

Our simulated power plant is composed of 4 continuous features: the pressure in the reactor's core, the temperature of the water in the reactor, the amount of water in the steam generator and the reactor's power. Furthermore, the power plant has four other discrete features that regard the reactor's rods: security rods, fuel rods, sustain rods, and regulatory rods. The first two have two levels: up and down. Instead, the latter two have three levels: up, medium, and down.

The reactor power linearly decreases over time for the effect of the de-potentiation of the fuel rods. Hence, the reactor's power depends on the values of the environment's features and whether nuclear fission is taking place. Moreover, the energy produced at each step is computed by dividing the reactor's power by 360, which is the power that a 1000MW reactor without power dispersion produces in 10 seconds (the amount of time we expected participants would need to act).

Actions to perform on the environment

The actions that the user can perform (12 in total) go from changing the position of the rods to adding water to the steam generator or skipping to the next step. All those actions change the value of 3 parameters, which correspond to the water's temperature in the core, the core's pressure, and the water level in the steam generator, respectively. The setting of the rods determines the entity of the feature updates; such updates are performed at the end of each step, right after the users' action. For example, if the safety rods are lowered in the reactor's core, the nuclear fission stops; thus, the temperature and pressure of the core decrease (unless they reach their initial values), and the water in the steam generator remains still. On the other hand, if nuclear fission occurs and the user lowers the regulatory rods, the fission accelerates. This acceleration consumes more water in the steam generator, raising the core's temperature and pressure more quickly and raising the reactor's power and electricity.

6.2.3 The robot's AI

Regarding the robot's AI model, we trained a deterministic decision tree (DT) using the Conservative Q-Improvement (CQI) learning algorithm (Roth et al., 2019), which allowed us to train the DT using a reinforcement learning (RL) strategy. CQI learns a policy in the form of a DT by splitting its current nodes only if it represents a policy improvement. Leaf node corresponds to abstract states and indicate the action to be taken, while branch nodes have two children and a splitting condition based on a feature of the state space. Over time, their algorithm creates branches by replacing existing leaf nodes if the final result represent an improved policy. In this sense, the algorithm is considered additive; while it is conservative in performing the splits (Roth et al., 2019).

Instead of extracting the DT from a more complex ML model (Vasilev et al., 2020; Xiong et al., 2017), we used this learning strategy to simplify the translation from the AI to the XAI without losing performance. The robot uses this expert DT to choose its action: it can perform each of the twelve actions based on the eight environment's features. We used (RL) because we had no data to train our ML model.

Starting from its root node, the DT is queried on each internal node - representing binary splits - to decide which sub-trees continue the descent. Each internal node regards a feature x_i and a value for that feature v_i: the left sub-tree contains instances with values of $x_i \leq v_i$, while the right sub-tree contains instances with values of $x_i > v_i$ (Buhrman and de Wolf, 2002).

The DT's leaf nodes represent actions; in the implementation of Roth et al. (2019), they are defined with an array containing the actions' expected Q-values: the greater Q-value is associated with the most valuable action. This way, the DT can be queried by users with both what- and why-questions. To answer a what-question, we only need to navigate the DT using the current values of the environment's features and present the resulting action to the user. To answer a why-question, we can present one of the features' values we encounter during the descent.

6.2.4 The robot's XAI

Since the AI model to explain is already transparent (Adadi and Berrada, 2018), we can directly exploit the DT to provide explanations by using one of the feature values we encounter during the descent of the tree.

As we have seen in Section 6.2.3, during the DT descent, we encounter a set of split nodes defined by a feature x_i and a value v_i; the direction of the descent tells us if the current

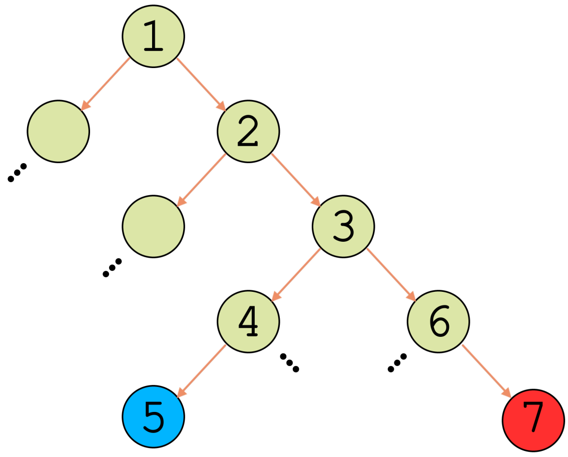


Figure 6.4 An example of DT where the leaf nodes 5 and 7 are the robot's suggestion and the predicted user's action, respectively. Let us assume that node 1 has already been used: the *classical* XAI selects node 2 for the explanations since it is the most unused relevant node. The *partner-aware* XAI selects instead node 3 because it represents a perfect contrastive explanation for fact 5 and foil 7.

scenario has a value of $x_i \leq v_i$ or $x_i > v_i$. Each of those inequalities can be used to provide an explanation that can help users to relate actions with specific values of the environment's features. In our case, an explanation for the action "add water to the steam generator" could be "because the water level in the steam generator is ≤ 25" (dangerously low).

Which feature to use above those we encounter during descent is a problem called *explanation selection*. In our case, we compared two explanation selection strategies. Classical approaches use the most relevant features (in terms of the Gini index, information gain, or other well-established measures (Stoffel and Raileanu, 2001)). We plan to compare a classical explanation strategy with a contrastive partner-aware one.

The *classical* XAI explains using only the AI outcomes and environment's states. In particular, it justifies the robot's suggestions using the most relevant features: the first ones in the DT's structure (see (Roth et al., 2019)). The system tries to give different explanations by taking track of the DT's node already used and preferring to use the others in decreasing the level of relevance.

On the other hand, the *partner-aware* XAI approach, through monitoring and scaffolding (Figure 6.1), takes into consideration the partner action indication and uses such indications to provide contrastive explanations: the fact (the outcome to explain) will be the robot's suggestion; in contrast, the foil (the expected outcome) will be the predicted users' action. We can say that the two explanation strategies used the same set of explanations (*e.g.*, *“because the power of the reactor is higher of 500 MWh”*, or *“because the fuel rods are up”*), and differed only in the explanation selection criteria. Figure 6.4 shows an example of these two approaches.

6.2.5 Measuring the explanations’ goodness

Assessing the model’s information power involves the interaction between non-expert users and the system itself. Thus, we need to collect measures for each rule and combine them to obtain the model’s information power. The general assessment steps are the following:

1. To quantify how many rules regard each feature and define a method for measuring the number of learned rules relative to each feature.
2. To quantify how much informative weight each feature has. For the sake of simplicity, we can assume that they have the same informative weight (equal to $\frac{1}{k}$, where k is the number of features). The features’ informative weights must respect the following rule: $\forall j \in \{1,...,k\}, \sum_j \gamma_j = 1$. The features’ informative weight describes how difficult it is to understand the rules regarding such features.
3. To measure the model’s informative power for each user, obtaining a measure (and a set of secondary descriptive measures) for each of them.
4. To average those measures obtaining the final results, the secondary descriptive measures could also have valuable meaning.

Hence, if $k \in \mathbb{N}$ is the number of the model’s features, $\gamma_j \in [0,1]$ is the informative weight of the feature j, $n^r_j \in \mathbb{N}$ is the number of rules regarding the feature j, $n^{lr(i)}_j \in \mathbb{N}$ is the number of rules regarding the feature j learned by the user i, and a_m is the *accuracy* of the AI model m, then the informative power of the model m for the user i is computed as follows:

$$IP_i(m) = a_m \sum_{j=1}^{k} \gamma_j \left(\frac{n^{lr(i)}_j}{n^r_j} \right) \in [0,1]. \tag{6.1}$$

Thus, if n^p is the number of users who took part in the assessment, the information power of the model m is

$$IP(m) = \frac{1}{n^p} \sum_{i=1}^{n^p} IP_i(m) \in [0,1]. \tag{6.2}$$

Apart from the number of rules regarding each feature, the most delicate aspect of the assessment regards the definition of the features' information weights. We suggest at least two ways to set them: to make them all equal or to define the weights using experimental data. Of course, which approach to follow is a task-related problem. A simple data-driven idea could be to set the features' information weight by normalising the number of interactions with the system that users need to understand those features.

Experimental measures

During our task, we plan to collect several quantitative measures to compute the model's information power as defined above:

- **M1**: Performance measures, such as the users' final score.
- **M2**: Measures of rules understanding, such as the number of task rules learned, the number of requests and interactions users needed to learn such rules, and the number of correct answers to the post-experiment test.
- **M3**: Measures of generalisation, such as the number of correct answers to *what-if* questions about the agent's decisions in particular contexts.

Moreover, we collect some subjective measures, like:

- **M4**: Satisfaction measures, such as users' satisfaction level about the explanations and the interaction.
- **M5**: Measures of agent perception, such as users' feelings towards the agent and perception of it.

6.3 Discussion

The assessment task we propose satisfies properties unique to the XAI research field and can be implemented in several HRI/HCI scenarios. Firstly, it focuses on the intended information power of the XAI system, which is the amount of information it can give to the user. Then, it defines the goodness of the XAI system as a function of its information power. Secondly, it allows for an objective and quantitative analysis of its impact on users' level of understanding.

The most critical requirement of our assessment task is user interaction. It allows for a two-way interaction with the users. In particular, it allows the users to query the system by asking what it would do in a specific situation and why. Consequently, the XAI system should be able to answer both what- and why-questions to exploit the full potential of the assessment task.

Another essential factor of this assessment framework is that one can easily generalise it. To reproduce the IP assessment in a different scenario, one needs the following:

- A decision-making task with the same characteristics as the one presented in Section 6.2.1.
- An expert AI and several non-expert users.
- The comparison approaches could be XAI algorithms or HCI/HRI dynamics.
- At least one quantitative measure about users' understanding of the task: if two or more, a method to compact them into a single measure is also needed.
- At least one quantitative measure about users' ability to generalise to unseen scenarios: if two or more, a method to compact them into a single measure is also needed.

Therefore, the decision-making task needs to be submitted to non-expert users giving them the objective of learning the functioning behind the task. During the task, participants should interact with an expert (X)AI model (under investigation) and ask it for help during the learning. Right after the learning phase, there should be an assessment phase in which the users need to prove their understanding of the task by performing it at their best. Finally, the experimenter should submit participants to the test to objectively measure their understanding of the task. An assessment task with those characteristics is flexible enough to test different AI models and XAI techniques as long as they allow user and system interaction.

In the next chapter, we present a comparison of two explanation strategies using such an assessment task. The first explanation strategy, which we called *classical*, selected the explanations in descending order of importance according to the DT structure. The second explanation strategy, which we called *adaptive*, provided contrastive partner-aware explanations using participants' action indications as the explanations' facts. Moreover, we performed experiments in both HCI, with a virtual speaking agent, and HRI, with the humanoid robot iCub. Then, we performed experiments without any (X)AI agent to compare such self-learning with the ones influenced by the artificial agents. We submitted participants the NPP task presented in Section 6.2.1, and considered the IP formula in its simplest form: we considered the task's rules equally important, and the accuracy of the AI model equal to 1. Subsequently, we submitted to participants a multiple-choices test containing a question for each task's rule to measure their understanding of the task. We considered the perfectly completed test of the participant i, equal to $IP_I(m) = 1$. Finally, we averaged the IP values among the participants to compare the information power of the two explanation strategies under consideration.

Chapter 7

Learning by doing with XAI in HCI and HRI

“Not making a decision is actually a decision. It’s the decision to stay the same.”
L. Terkeurst, Best Yes

In this section, we show the results of the experiments in which we used the assessment task we presented in the previous section. In particular, we present and compare the results of two groups of participants who interacted with a virtual artificial agent and the humanoid robot iCub, respectively. In the next section, these findings will be compared with those we obtained with a third group of participants who interacted with no artificial agent.

7.1 Methods

The experiment was composed of five phases. During the first phase, we asked participants to complete the Big 5 personality traits questionnaire (Gosling et al., 2003) and several items about their previous experiences with computers or the robot iCub. At the end of such a questionnaire, we asked them to describe the functioning of a nuclear power plant (NPP). We used their open-ended answers to code their level of understanding of NPPs.

During the second phase - the *training* - participants had to perform the task for 30 minutes. Before starting it, we instructed them on how to use the control panel and on their goal for this phase: to try to understand the task’s objectives and rules as much as possible. Depending on the experimental condition, participants could interact with an explainable AI agent (the computer or the robot iCub) and ask it what- and why-questions. In each step, participants had to indicate their action first (*indication of action*), then they could ask

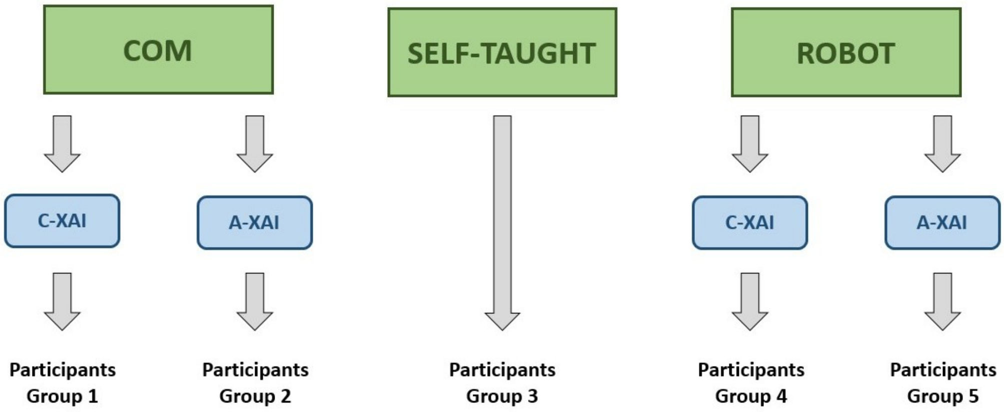


Figure 7.1 The experimental design we used for the NPP experiment. There were three macro-groups: COM (i.e., with the computer), Self-taught, and Robot. We split participants from the COM and Robot groups in two experimental conditions (C-XAI and A-XAI), while we investigated only one condition with the Self-taught group.

questions to the artificial agent and/or confirm their action. As long as they did not confirm the action, they did not actually perform it and move to the next step.

The third phase - the *assessment* - replicated the same task but without the help of any AI agent. Before starting, we instructed participants that they had to perform at their best the same task they learned about in the previous phase. Hence, they had to fulfil the objectives and follow the rules they had learned. The assessment phase lasted for 10 minutes.

The fourth phase corresponded to a post-experiment questionnaire. It contained the same items about the agent (the virtual one or iCub, depending on the experimental groups) we submitted during the pre-experiment questionnaire. This way, we could compare the pre- and post-interaction participants' impressions of the agents. All questionnaires' items can be found in the appendix of this thesis.

During the last phase, participants had to complete a test to assess their understanding of the task. The test was divided into two parts. The first part contained two open-ended questions: the first asked to describe the functioning of the NPP, while the second asked to describe the functioning of the agent participants interacted with. The second part of the test contained 34 multiple-choice questions (4 answers per question): two questions about the task's objectives, four about the rods' purposes, nine about the task's rules, and thirteen in which participants had to select the effect of actions taken in particular scenarios.

We performed experiments with two groups of participants ($n = 21$ each in total): COM and Robot. The former interacted with a virtual artificial agent (a voice on the computer), while the latter interacted with the humanoid robot iCub. We performed between-subject user studies for each of such groups by manipulating the agent's explanation selection strategy.

For the C-XAI (Classical XAI) group, the agents selected explanations in a decreasing importance order: the DT that the agents used to provide both suggestions and explanations has been built in order to have the most important features on top and the less important ones towards the bottom (Roth et al., 2019). On the other hand, the A-XAI (Adaptive XAI) received contrastive explanations based on participants' indication of action: the agents selected the node of the DT representing the reason why they would perform their suggested action rather than the participants' indication of action. Figure 6.4 illustrates the difference between the two strategies.

Finally, we performed the same experiments without any artificial agent. Participants in the Self-taught group ($n = 10$ in total) performed the task without any help - thus, without the chance of asking what- and why-questions - and took the same post-experiment test as the other experimental groups. Figure 7.1 shows the experimental groups and conditions and how we split them.

We split the results in two main subsections regarding to the influence of the virtual agent and the influence of the humanoid social robot. Regarding the experiment with the virtual agent, we expected that:

- H1: adaptive explanations would elicit more accurate participants' mental models about the task because we expected that the contrastive nature of such explanations would help participants understand the task's rules better and faster.

- H2: no particular differences between the two conditions regarding the behavioural measures, such as number of actions performed, or moving time.

- H3: no effects of the explanation strategies on the agent's persuasiveness because the interaction lacked the social component.

- H4: participants' personality traits would not affect the interaction with the agent.

On the other hand, regarding the experiment with the social robot, we expected that:

- H5: adaptive explanations would elicit more accurate participants' mental models about the task than for the COM group.

- H6: no particular differences between the two conditions regarding the behavioural measures.

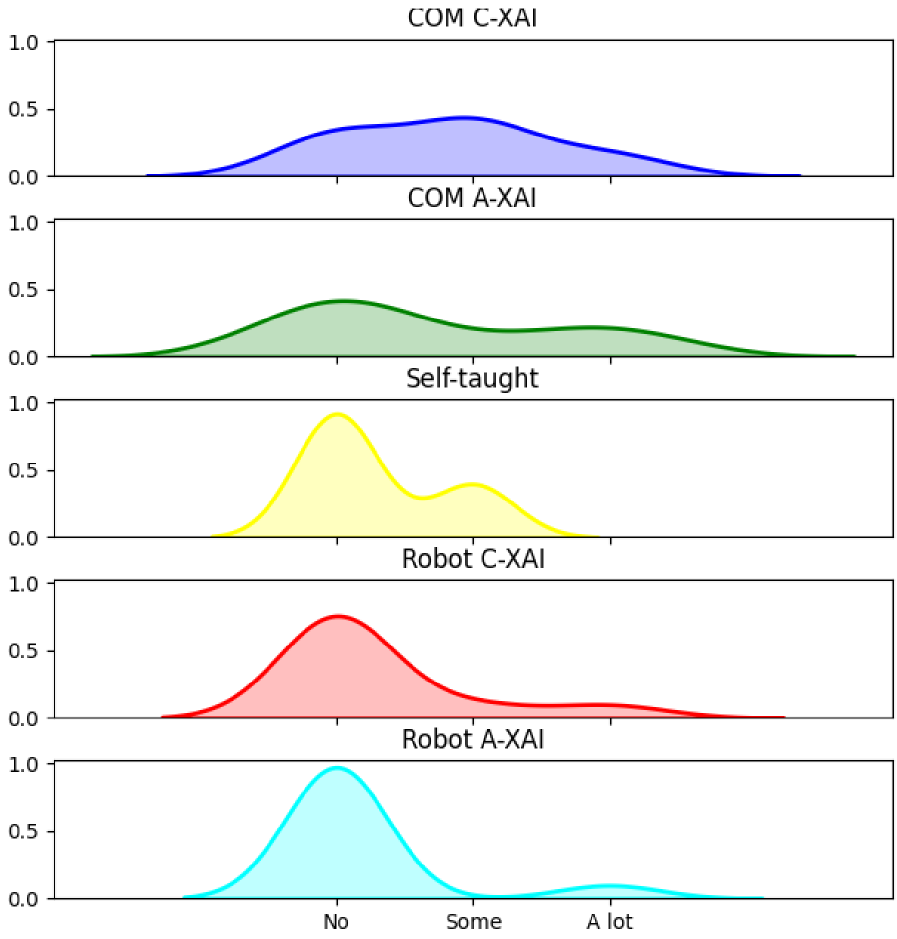


Figure 7.2 Distribution of the participants' knowledge about the functioning of nuclear power plants before the experiment. We classified them into three levels of knowledge (No, Some, A lot) by coding their open-ended questions to the pre-experiment questionnaire.

- H7: strong effects of the explanation strategies on the agent's persuasiveness because we inserted the social component in the interaction.
- H8: participants' personality traits would affect the interaction with the agent.

7.2 Results

7.2.1 A priori knowledge of the task

Since we used a between-subject approach, we needed to be sure that the different groups had comparable starting conditions, i.e. that they started with similar task knowledge. To

	COM				**Robot**			
Explanandum	t	p	μ_C	μ_A	t	p	μ_C	μ_A
Temperature	50.86	<.001*	22%	7%	3.76	<.001*	24%	15%
Pressure	11.58	.001*	28%	14%	3.74	<.001*	22%	10%
Water steam gen.	13.43	<.001*	19%	30%	.81	.42	24%	27%
Power	1.46	.23	16%	17%	1.17	.24	17%	18%
Safety rods	13.87	<.001*	7%	18%	-2.2	.03*	5%	14%
Regulatory rods	10.45	.002*	1%	3%	.06	.94	0%	0%
Sustain rods	.77	.38	8%	9%	-1.8	.04*	8%	15%
Fuel rods	2.92	.09	1%	2%	.1	.91	1%	1%

Table 7.1 Comparison of the occurrences (in percentage) of the explanandum between the experimental conditions: we signed with an * those comparisons that were significantly different (independent samples t-test). Five out of eight comparisons were significantly different in the COM group, and four out of eight in the Robot one.

measure participants' knowledge about the functioning of NPPs, we asked them to describe such functioning in an open-ended question in the pre-experiment questionnaire. Then, we coded their answers in three levels of understanding (No, Some, and A lot of knowledge) and performed a χ^2 test to investigate whether there were differences in the distribution of participants' prior levels of understanding among the experimental groups. We found no significant differences between such distributions (χ^2 test: $\chi^2(8) = 14.1$, $p = .078$). The distribution of participants' level of prior knowledge is shown in Figure 7.2 and are all comparable between each other meaning that the experimental groups started with similar prior knowledge about the general functioning of NPPs.

7.2.2 Quantitative differences between classical and adaptive explanations

As already mentioned, we manipulated our experiments on the type of explanations the artificial agent provided. For one group of participants, we provided *classical* explanations (C-XAI), while for the other group we provided contrastive *adaptive* explanations (A-XAI). Since the difference between the two explanation strategies resided in the explanations selection, we counted (in percentage) the number of explanations' subjects (the *explanandum*) used by the artificial agent during the training phase; we compared such percentages among the experimental groups to check whether different explanation strategies brought to different explanandum.

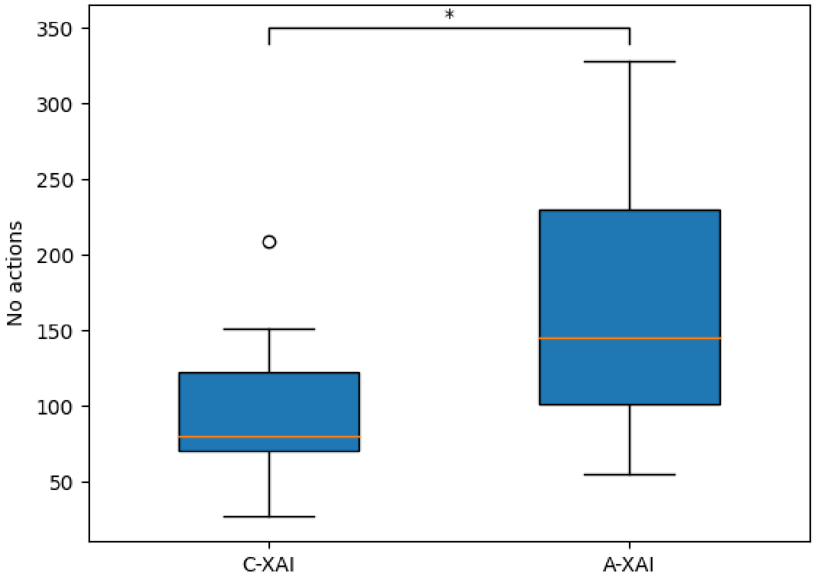


Figure 7.3 Number of actions participants in the COM group performed during training. The * represents a statistically significant difference with *p-value* of .039 (independent samples t-test).

We performed an independent samples t-test on each possible explanandum (the environment's features) between the C-XAI and A-XAI groups for both the COM and Robot groups. Regarding the COM group, we found significant differences in five out of eight comparisons between explanandums' percentages. Instead, regarding the Robot group, we found significant differences in four out of eight comparisons. Table 7.1 shows all the comparisons with their statistics.

Moreover, we investigated whether there were differences also in the questions asking frequencies between the conditions. Hence, we performed a χ^2 test on the questions' frequencies. Such tests showed no significant differences between the groups about both the what- and why-questions.

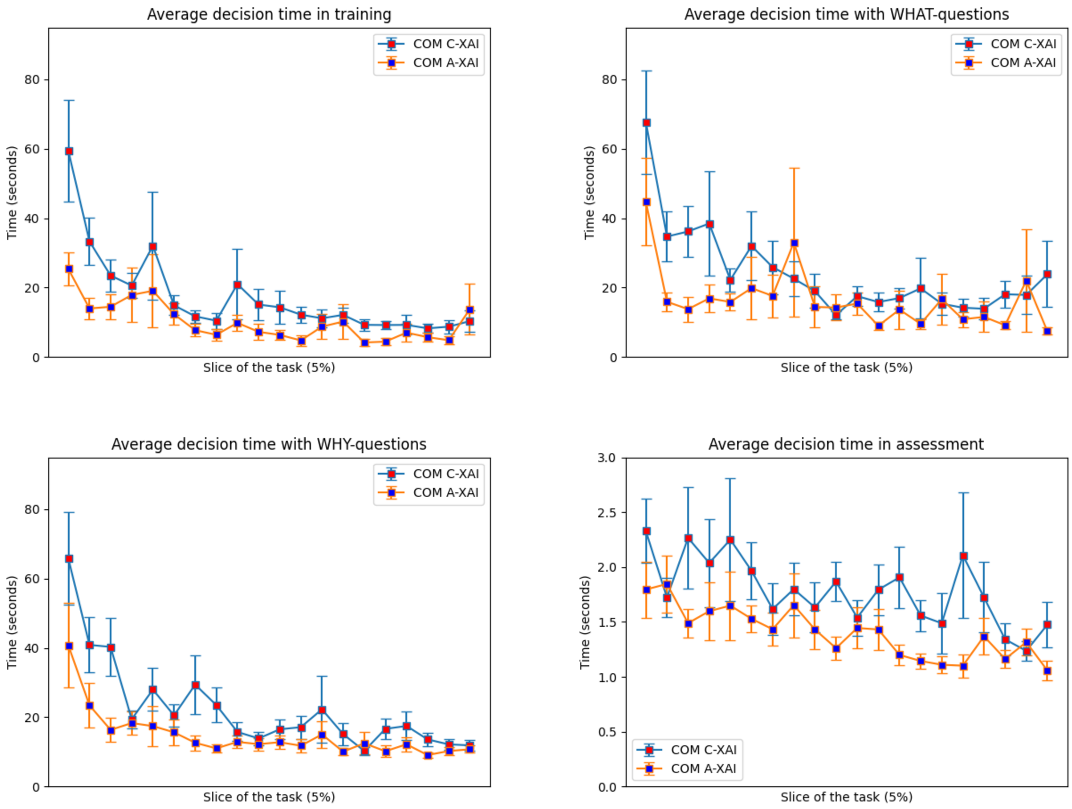


Figure 7.4 Average and std error of participants' decision times (COM group). Each point of the x-axis represents 5% of the training phase: it has to be read from the left to the right. The plot on the up-left side shows the average decision time of all steps; the one on the up-right side regarded those where participants asked what-questions; the plot on the bottom-left side regarded those where participants asked why-questions. The one on the bottom-right side shows the assessment decision time and uses a different scale for the y-axes to improve readability. The plots show that participants in the A-XAI group were faster in performing the actions than those in the C-XAI one.

7.2.3 The influence of a virtual artificial agent

In this section, we show the results related to the COM group.

We found that participants who received adaptive explanations performed significantly more actions than those who received classical explanations (independent samples t-test: $t = 2.21$, $p = .039$). Figure 7.3 shows a box plot of participants' actions during the training phase. However, we did not find differences in the behavioural measures regarding the assessment phase between the two groups.

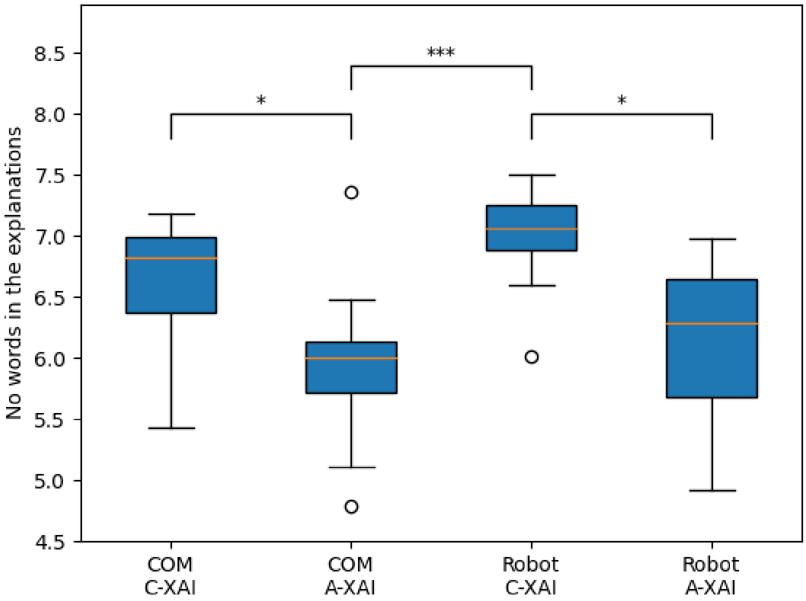


Figure 7.5 Explanations' length measured in the number of words used by the artificial agents. The * refers to a significant difference with $p-value < .5$, and the *** refers to a strong significant statistical difference with $p-value < .001$ (independent samples t-test).

This result is explained by the significant difference between the average moving time of the two groups (independent samples t-test: $t = 2.3$, $p = .02$), as shown in Figure 7.4. We also found that the decision time of the steps in which participants asked what- and why questions were significantly different between the C-XAI and A-XAI groups (independent samples t-test: $t = 2.19$, $p = .03$, and $t = 2.3$, $p = .02$, respectively). Finally, we found that the decision times during the assessment phase differed significantly between the two groups (independent samples t-test: $t = 4.4$, $p < .001$), with the A-XAI group moving significantly faster than the C-XAI one.

Since participants performed different numbers of moves, we aggregated those in slices (on the *x-axis* in Figure 7.4) representing pieces of 5% of the task to compare their moves' decision times. Thus, we compared the participants' corresponding slices to have measures of similarity and easily visualised their moving times.

However, we found that participants in the A-XAI group received explanations less verbose than those in the C-XAI group (independent samples t-test: $t = 2.31$, $p = .03$) as shown in Figure 7.5. We found no correlations between participants' personality traits, assessment behavioural measures, or persuasiveness. To summarise these findings, we can say

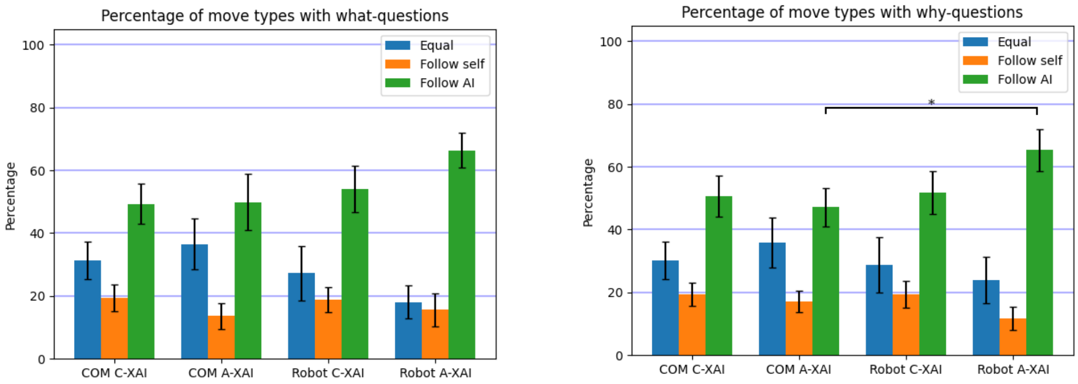


Figure 7.6 Average and std error of participants' move types during the training phase of both the COM and Robot groups. The plot on the left refers to the step in which participants asked what-questions, while the one on the right refers to the step in which they also asked why-questions. The difference reported with the * refers to a *p-value* of .038 (independent samples t-test). We found that, with the A-XAI, the percentage of follow AI moves of participants in the Robot group was significantly higher than those of participants in the COM one.

that participants who interacted with the virtual agent moved faster and were more resolute with the adaptive explanations than with the classical ones. This brought who received the former to perform a higher number of actions than the other group of participants, since the task was time-bounded.

7.2.4 The influence of a humanoid social robot

In this section, we show the results related to the Robot group. Contrary to the COM group, we found no behavioural difference between the two groups that interacted with the robot. This is stated both for the training and the assessment phases. We found that participants who interacted with the iCub robot performed a comparable number of actions during the training, regardless of the experimental group to which they belonged. Indeed, contrary to the COM group, we found no significant differences regarding the participants' decision time between the C-XAI and A-XAI groups.

However, we found that participants in the A-XAI group received explanations less verbose than those in the C-XAI group (independent samples t-test: $t = 2.31$, $p = .03$) as shown in Figure 7.5.

As for the previous group, we found no correlations between the participants' personality traits and their behavioural measures during training. However, regarding the assessment phase, we found that the amount of energy produced negatively correlated with participants' positive agency (Pearson's $r = -.455$, $p = .038$) and positively with their negative agency (Pearson's $r = .484$, $p = .026$). Moreover, we found that also the number of anomalies negatively correlated with participants' positive agency (Pearson's $r = -.449$, $p = .041$).

7.2.5 Comparisons between the virtual agent and the humanoid robot

In this section, we compared the results of the COM and Robot groups. To measure the influence that such artificial agents had on participants, we collected three types of participants' moves for each step of the training phase:

- *Equal*: participants' first indication and the agents' action were the same from the beginning.
- *Follow self*: participants' first indication and the agents' action differed, but they chose to confirm their initial indication.
- *Follow AI*: participants' first indication and the agents' action differed, and they followed the agents' advice.

To compare the move types of participants from different experimental groups, we averaged the percentages of their move types and tested those through independent samples t-tests. Regarding the what-questions, Figure 7.6 (left side), we found no significant differences in this regard. On the other hand, regarding the why-questions, Figure 7.6 (right side), we found a significant difference in the adaptive explanations between the COM and Robot group (independent samples t-test: $t = 2.226$, $p = .038$ ($\mu = 47.08$, $\sigma_M = 6.1$) for the COM group, and $t = 2.226$, $p = .038$ ($\mu = 65.37$, $\sigma_M = 6.68$) for the Robot group). Hence, we had that the adaptive explanations made the robot more persuasive than the artificial agent when they justified their suggestions. Finally, Table 7.2 shows the significant correlations (Pearson) we found between behavioural measures and move types in both COM and Robot groups.

When interacting with the robot, the two explanation strategies did not caused any behavioural changes participants. However, the adaptive explanations resulted in a higher robot's persuasiveness; thus, participants relied more on the robot's suggestions when justified with adaptive explanations than with classical ones. To summarise those findings, we can say that adaptive explanations automatised participants behaviour when interacting with the virtual agent, while they raised participants' tendency to follow the agent's suggestions when interacting with the social robot.

	COM			**Robot**		
	Equal	Follow self	Follow AI	Equal	Follow self	Follow AI
Tot. energy	r = .546 p = .011	-	r = -.513 p = .017	-	-	r = -.559 p = .008
Tot. actions	r = .644 p = .002	-	r = -.501 p = .021	-	-	r = -.747 p <.001
Tot. anomalies	-	r = .581 p = .006	-	-	r = .593 p = .005	r = -.558 p = .009
Anomaly rate	-	r = .492 p = .024	-	-	r = .475 p = .03	-
Tot. critic steps	r = .734 p <.001	-	-	-	-	r = -.658 p = .001
Critic rate	r = .546 p = .010	r = -.572 p = .007	-	-	r = -.484 p = .026	-
Tot. what-questions	r = .725 p <.001	r = -582 p = .006	-	r = .516 p = .017	r = -.563 p = .008	-
What-questions rate	-	r = -.761 p <.001	r = .641 p = .002	-	r = -.681 p <.001	r = .731 p <.001
Tot. why-questions	r = .676 p <.001	r = -.74 p <.001	-	-	r = -.663 p = .001	r = .473 p = .03
Why-questions rate	-	r = -.647 p = .002	r = .601 p = .004	-	r = -.605 p = .004	r = .768 p <.001
Tot. energy	r = .452 p = .04	-	-	-	-	-
Tot. actions	r = .667 p <.001	-	-	r = .516 p = .017	-	-
Anomaly rate	r = -.47 p = .031	-	-	r = -.489 p = .024	-	-
Tot. critic steps	r = .624 p = .002	-	-	r = .647 p = .002	-	-
Critic rate	-	-	-	r = .461 p = .036	-	-

Table 7.2 Significant correlation (Pearson) between move types and behavioural measures. The values above the line refer to the training phase, while those below the line refer to the assessment one.

7.3 Discussion

Regarding the influence of the virtual agent on participants' behavior and learning, we expected that (H1) adaptive explanations would elicit more accurate participants' mental models about the task, (H2) to find no particular differences between the two conditions regarding the behavioural measures, (H3) to find no effects of the explanation strategies on the agent's persuasiveness , and that (H4) participants' personality traits would not affect the interaction with the agent. On the other hand, regarding the influence of the humanoid social robot, we expected that (H5) adaptive explanations would elicit more accurate participants' mental models about the task, (H6) to find no particular differences between the two conditions regarding the behavioural measures, (H7) to find strong effects of the explanation strategies on the agent's persuasiveness, and that (H8) participants' personality traits would affect the interaction with the agent.

We analysed participants prior competences about nuclear power plants to ensure the similarity of the initial conditions of every group involved in the experiments. To do so, we asked participants to describe the functioning of an NPP (whose management represented the experimental task) during the pre-experiment questionnaire. Consequently, we coded such open-ended answers in three levels of knowledge and checked if there were differences in the distribution of the knowledge between the groups. Although such distributions were comparable, we can note a difference in those referring to the COM group and the other two: Self-taught and Robot. This visual difference can be explained by the origin of the participants belonging to the groups: those in the COM lived in Germany, while those in the other two lived in Italy. We hypothesised that German people have a higher awareness of nuclear energy, because it has been used more recently compared to Italy; this brought to a slightly better knowledge about the functioning of NPPs.

Our between-subject factor was the explanation selection strategy. Both the artificial agents provided two kind explanations, which we called *classical* (C-XAI) and *adaptive* (A-XAI). The first strategy selected the "most important feature" (in descending order) that the AI model used for its classifications, while the other one provided contrastive explanations using participants' indications of actions.

We checked whether those explanation selection strategies brought different occurrences of the explanandum (the subject of the explanation). We observed that this was the case for more or less half of the possible explanandum, especially for those with higher occurrences. However, it has to be noted that the explanations, thus the explanandum, strictly depended on the participants' behaviour during the learning. Hence, such dependency could also explain the differences discussed here.

Adaptive explanations caused participants who interacted with the COM virtual agent to move faster than those who received classical ones. We observed such an effect during both the training and the assessment phases; thus, the explanation strategy affected participants' behavioural measures, rejecting our hypothesis H2. Indeed, during training, the A-XAI group moved faster than the C-XAI one when they asked what- and why-questions, but the moving time between the two groups was comparable when no question-asking was involved. For this reason, we can say that the difference in the moving time was mainly due to the explanation type that participants received. However, we found that classical explanations were more verbose than adaptive ones: this could have contributed to such differences. In our opinion, the difference in explanations' length was not enough because they differ of one word in average: it is not enough to claim that the explanations' length was responsible for participants moving times.

Moving faster brought participants in the A-XAI group to perform more actions and produce more energy in training and assessment than in the C-XAI group. Despite those differences, the two explanation styles produced comparable exploration strategies in training (*e.g.*, number of anomalies in training) and level of understanding of the task (*e.g.*, results at the test), rejecting our hypothesis H1. Hence, we can say that the two explanation strategies brought only behavioural differences in the HCI context. Moreover, we found no effects of participants' personality traits on the interaction with the virtual agent, confirming our hypothesis H4.

On the other hand, regarding the Robot group, we found no behavioural differences brought by the two explanation strategies, confirming our hypothesis H6. Indeed, we observed a comparable number of actions performed, anomalies produced and moving times between the C-XAI and A-XAI groups, both in training and assessment.

However, adaptive explanations influenced participants more with the humanoid robot than with the virtual agent, confirming H3 and partially confirming our hypothesis H7. Indeed, the adaptive robot persuaded participants to opt for its actions (rather than their first indication) more than the virtual agent did. Hence, we can say that adaptive explanations, on the one hand, empowered participants, making them more confident while interacting with the virtual agent; on the other hand, they persuaded more participants who interacted with the humanoid robot.

Nonetheless, the two explanation strategies proved to be similarly informative also for the Robot group, rejecting our hypothesis H5.

Contrary to what we expected, we found no correlations between the personality traits and behavioural measures of participants who interacted with the robot during training but only during the assessment phase. Thus, our hypothesis H8 has been partially confirmed. Indeed, the amount of energy produced in assessment positively correlated with participants' negative agency and negatively with their positive agency. Thus, the higher their negative agency and the lower their positive agency, the higher their performance in assessment. We speculate that the participants who tended to automatise their behaviour - thus, not choosing consciously and intentionally their moves - had an advantage regarding their performance in assessment. The further negative correlation between the number of anomalies in assessment and participants' positive agency supports our speculation about behaviour automatism.

The correlations between participants' behavioural measures and move types tell us how the artificial agents influenced both the training and assessment. Regarding the training phase, we can see that the number of anomalies positively correlated with the percentage of follow-self moves with both agents. Moreover, the former negatively correlated with the percentage of follow-AI moves only with the robot. It seems reasonable that less non-expert users followed the expert agent's suggestions the more they made errors; similarly, the opposite seems reasonable as well, but turned out to hold only with the iCub robot.

Similar considerations can be made regarding the critic rate (the portion of task where the reactor was functioning), which positively correlated with the equal moves and negatively with the follow-self ones with the COM but only negatively correlated with the follow-self moves with the robot. Following the expert agent would enhance one's critic rate and vice versa. However, reproducing the same agent's actions helped in this regard only with the virtual agent.

Regarding assessment, we can see negative correlations between participants' anomaly rate and the percentage of equal moves for both COM and Robot groups. On the other hand, we observed the opposite regarding participants' number of critic steps. It positively correlated with the equal moves for both groups. These results tell us that participants of both the groups performed better in assessment when they could replicate the artificial agents' moves.

The main limitation of this study regard the high variability of participants' behavior during both training and assessment phases, the use of self-reported questionnaires, and possible biases that we might have inserted into the final test. Since they had no limitation in actions, participants showed a high variability in performing their actions; thus, we could aggregate and analyse them in percentage. Moreover, to limit possible biases of self-reported questionnaires we used the same precautions used in the previous studies. Finally, we designed the final test considering the design principles from existing literature in multiple choice questions making.

In this chapter, we discussed the results of the user study where we implemented the information power assessment task presented in the previous one. We used the task to compare the informativeness of two explanation strategies: a classical strategy, which provided explanations based on their relevance, and a partner-aware one, which provided contrastive explanations based on users' intentions of action. Our assessment task did not highlight particular differences in the informativeness of such strategies. However, they brought to different users' behavior depending on the artificial agent they interacted with. Indeed, we let participants interact with a virtual speaking agent or a humanoid social robot. We hypothesise that the difference between the two explanation strategies was too thin to bring to differences in participants' mental models about the task, and that the expert agents had a more intrusive influence on participants' behavior. For this reason, we opted for a further study where people have to perform the learning-dy-doing task on their own, without the help of any artificial agent. The reader can find the results of this study in the next chapter.

Chapter 8

Learning by doing without XAI

"The straight line, a respectable optical illusion which ruins many a man."
V. Hugo, Les Misérables

In this chapter, we present the results of the experiment which shares the methods of the one presented in the previous chapter. Thus, we highlight only the differences between the two. In the previous study, participants could interact with an artificial agent (virtual or robotic) to ask for help during the training phase. Contrary, the group of participants we consider in this chapter could not interact with any artificial agent while learning the task. In the following sections, we present a comparison between all groups of participants.

8.1 Results

8.1.1 About the self-learning

In this section, we present the results regarding the Self-taught group. Regarding the experiment without artificial agents, we expect that (H9) the lack of explanations would penalise Self-taught participants' mental models about the task with respect to those in the COM and Robot groups. On the other hand, we expect to observe (H10) participants' personality traits influencing their behaviour during training and assessment since they had to perform them independently.

This group of participants performed the same task without interacting with any artificial agents. Thus, they could not ask anybody for help with what- and why-questions. As a result, they had to learn on their own how the NPP worked.

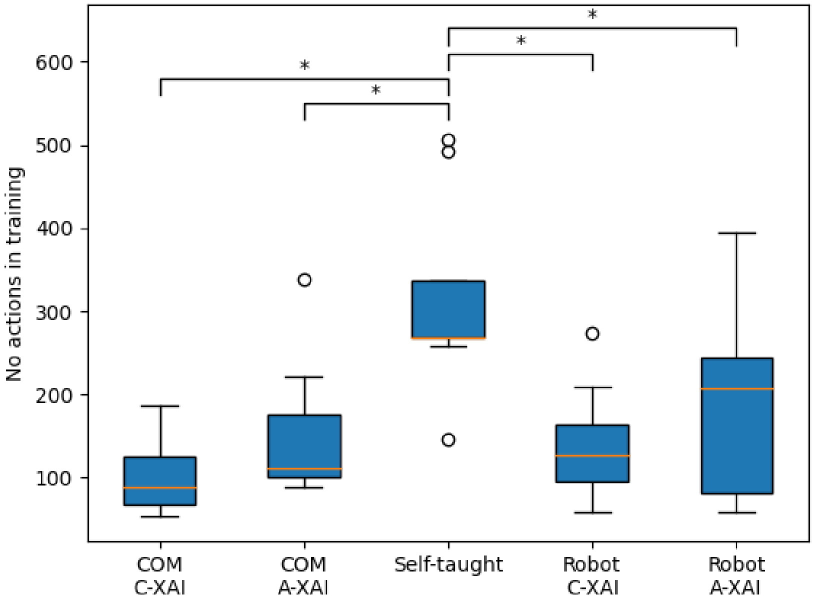


Figure 8.1 Number of actions performed during the training phase by all the groups. The * refers to strong statistical significance (ANOVA test with Bonferroni correction).

8.1.2 Comparisons with the assisted-learning

Behavioural measures

We found that there was an effect of the experimental condition on the number of actions performed in training among the experimental groups (ANOVA $F(4) = 8.66$, $p < .001$). Through a post hoc test with Bonferroni correction we found that the Self-taught group performed more actions in training than all the other groups (Figure 8.1): $t = 3.654$, $p = .006$ with COM A-XAI group; $t = 5.556$, $p < .001$ with COM C-XAI group; $t = 3.24$, $p = .022$ with the Robot A-XAI group; and $t = 4.459$, $p < .001$ with the Robot C-XAI group.

Similarly, there was an effect of the experimental condition on the energy produced during the training between the experimental groups (ANOVA $F(4) = 10.9$, $p < .001$). Through a post hoc test with Bonferroni correction, we found that the Self-taught group produced more energy than all the other groups : $t = 4.684$, $p < .001$ with COM A-XAI group; $t = 5.943$, $p < .001$ with COM C-XAI group; $t = 3.992$, $p = .002$ with the Robot A-XAI group; and $t = 5.344$, $p < .001$ with the Robot C-XAI group.

We also found a statistically significant difference regarding the number of anomalies in training (ANOVA $F(4) = 6.86$, $p < .001$). Through a post hoc test with Bonferroni

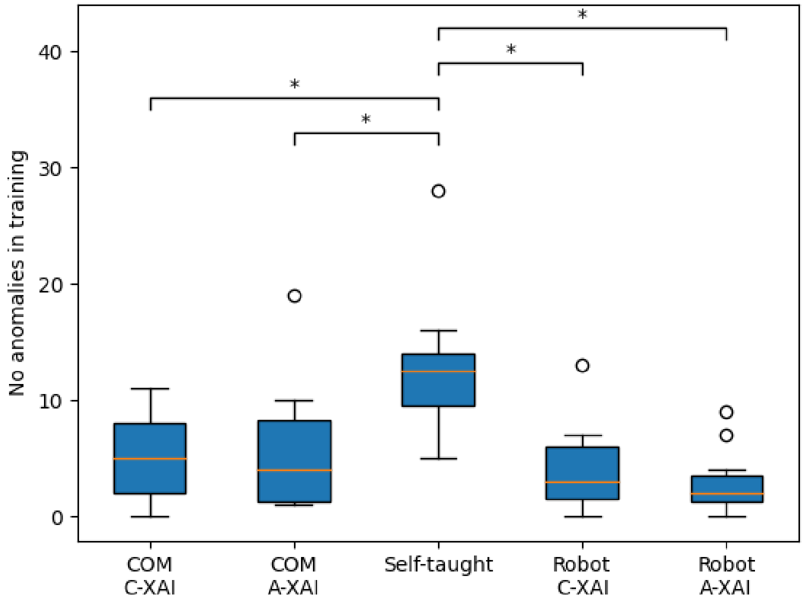


Figure 8.2 Number of anomalies produced during the training phase by all the groups. The * refers to strong statistical significance (ANOVA test with Bonferroni correction). Self-taught participants produced more anomalies than those in the other groups.

correction, we found that the Self-taught group produced more anomalies than all the other groups (Figure 8.2): $t = 3.345$, $p = .016$ with COM A-XAI group; $t = 3.766$, $p = .005$ with COM C-XAI group; $t = 4.255$, $p < .001$ with the Robot A-XAI group; and $t = 4.683$, $p < .001$ with the Robot C-XAI group.

Finally, we found an effect of the experimental condition on the number of critic steps in training (ANOVA $F(4) = 6.52$, $p < .001$). Through a post hoc test with Bonferroni correction, we found that the Self-taught group performed more critic steps than the COM C-XAI group ($t = 3.293$, $p = .015$), the COM A-XAI group ($t = 4.817$, $p < .001$), and the Robot C-XAI one ($t = 3.765$, $p = .004$).

These results are explained by the less decision time participants in the Self-taught group had with respect to the COM and Robot groups (Figure 8.3) (independent samples t-test: $t = -3.62$, $p < .001$, and $t = -4.12$, $p < .001$, respectively).

Post-experiment test

We found an effect of the experimental condition on the percentage of correct answers to the whole test when considering the Self-taught group (ANOVA $F(4) = 3.99$, $p = .007$).

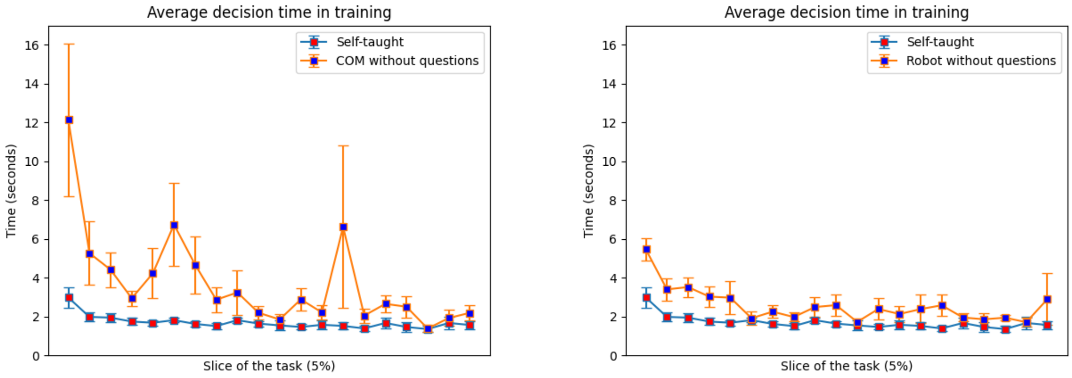


Figure 8.3 Average and std error of participants' decision times during training. Each point of the x-axis represents 5% of the training phase: it has to be read from the left to the right. The plot on the left shows the comparison between the decision times of the Self-taught group and the COM group when they asked no questions. The plot on the right shows the comparison between the decision times of the Self-taught group and the Robot group when they asked no questions.

Through a post hoc test with Bonferroni correction, we found that participants belonging to the Self-taught group outperformed the COM C-XAI group ($t = 3.422$, $p = .013$), the COM A-XAI group ($t = 3.28$, $p = .02$), and the Robot A-XAI one ($t = 3.065$, $p = .036$) at the final test.

None of the test's sections showed significant differences between the conditions, with the exception of the *scenarios* (ANOVA $F(4) = 6.37$, $p < .001$). Through a post hoc test with Bonferroni correction, we found that the Self-taught participants outperformed at the test's questions with the scenarios the COM C-XAI group ($t = 4.212$, $p = .001$), the COM A-XAI group ($t = 4.129$, $p = .001$), and the Robot A-XAI one ($t = 3.919$, $p = .003$). Figure 8.4 shows the distributions of correct answers to the whole test (left side) and at the scenarios part (right side).

We found no interesting correlations between the Self-taught participants' behavioural measures and their personality traits, confirming our hypothesis H10.

8.2 Discussion

As we expected, we observed that Self-taught participants performed more actions during training than those who interacted with both the artificial agents. This result is easily explainable by the greater availability of time they had with respect to those participants who

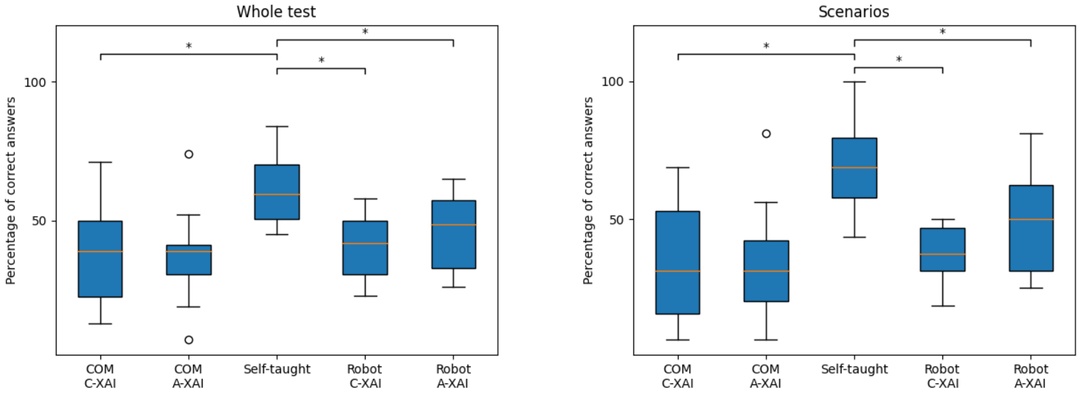


Figure 8.4 Distribution of the participants' correct answers to the whole test (left) and to the scenarios questions (right). The * refers to *p-values* between .02 and .001. We can see that Self-taught participants outperformed at test all those of the other experimental groups.

had to wait for the answers to their questions. Hence, Self-taught participants moved faster than the others, also when the latter did not ask any questions to the artificial agents.

Consequently, we found that they produced more energy and performed more critic steps than those who interacted with the artificial agents. This is reasonable because we found that both those measures positively correlated with the number of actions performed. More interestingly, we observed that Self-taught participants produced more anomalies than the other groups.

In our opinion, the number of anomalies during training is a good estimator of participants' degree of exploration. Indeed, exploring the environment was crucial to understanding its functioning and achieving good test results. We observed that Self-taught participants outperformed those who interacted with the artificial agents at the post-experiment test, especially regarding the scenarios questions, rejecting our hypothesis H9. Those questions presented several environmental scenarios (with textual descriptions and pictures) and asked what would happen if one performed a specific action. We found significant differences between the Self-taught group and all the others but the COM A-XAI one. However, we can see that this was because of an outlier in such a group, since it is more than two standard deviations away from the mean (Figure 8.4): if we remove it, we can easily achieve a significant effect.

Since the number of anomalies in training is a good indicator of the exploration, we can say that Self-taught participants explored the environment more than the others. We speculate that this was so because no agent influenced them; thus, they felt free to learn how the environment worked adequately without the pressure to avoid producing anomalies in front of an expert artificial agent. We can further hypothesise that self-taught participants did not suffer the harmful effect of the *anchoring* and *automation* biases (Vered et al., 2023), compared to those who interacted with the artificial agents. Those are the most frequent cognitive bias - heuristics that the human brain produced to facilitate and speed up the decision-making processes - in human-AI collaboration, by which people remains anchored to previous or others' ideas and suggestions. However, to claim the intervention of cognitive biases in participants' decision-making needs deeper investigation.

Considering that we found no behavioural differences between Self-taught participants and the others regarding the assessment phase confirms our hypothesis that interacting with expert explainable artificial agents influenced the training, but did not necessarily improve the training efficacy. Moreover, the Follow AI moves were the most present move type in all conditions involving an interactive agent. We think participants in the COM and Robot groups limited themselves by asking many questions and following the agents' suggestions uncritically. Hence, we need to reflect on how to deal with collaborative robots and AIs to ensure to not limit people's learning, since they seem to over-rely on such artificial agents.

We consider the fact that Self-taught participants explored more the environment a positive feature because this brought them in building a more consistent understanding of the task. On the other hand, participants in the COM and Robot groups seemed to be heavily influenced by the expert agent they interacted with. Such an influence brought to automation bias and over-reliance. Several strategies have been found in the literature to mitigate the automation bias resulting from the artificial agents' influence. Those comprehend cognitive forcing, which are mechanisms to help people re-think about their decision making. However, implementing cognitive forcing mechanisms in the setting of our IP assessment task may be difficult and represents a possible future work to improve people's performance during XAI-assistend learning-by-doing tasks.

Part V

Discussion

Chapter 9

Conclusions

"It is not our part to master all the tides of the world, but to do what is in us for the succour of those years wherein we are set, uprooting the evil in the fields that we know, so that those who live after may have clean earth to till. What weather they shall have is not ours to rule."
Gandalf - J.R.R. Tolkien, The Lord of the Rings

The act of providing explanations manifests in various forms and serves diverse purposes. Throughout this thesis, we predominantly treated explanations as justifications or motivations behind the robot's actions. Hence, we identified explainability as a crucial future challenge in the realm of social robotics research, with particular relevance to human-robot collaboration contexts. The integration of explainable autonomous robots holds promise for enhancing collaboration by aligning with human mental models, facilitating a more intuitive understanding of their behaviour.

Following a comprehensive introduction to the field of explainability and explainable artificial intelligence (XAI), we discussed the potentiality of XAI approaches that consider the explanation partner. We compared partner-aware XAI approaches to more classical ones in different scenarios. Our investigation included the influence of an explainable social robot during collaborative decision-making tasks and discussed the enhancement of explanations' persuasiveness through the exploitation of the human-robot common ground. Finally, we proposed an assessment task to objectively and quantitatively measure the informativeness of explanations for non-expert users while learning new tasks. Additionally, we investigated the influence of different artificial agents on participants' learning and discussed the urgency to rethink the design of tutor agents.

The primary objective of this work was to address the explainability problem through a Human-Robot Interaction (HRI) lens. We started by formulating a theoretical framework for XAI in HRI, which stressed the social-dialogical nature of robots' explainability. Hence, the implementation of the framework in different scenarios facilitated the investigation of partner-aware XAI during human-robot collaboration. In particular, we inquired whether explanations considering the partner influenced participants more than classical approaches. Lastly, we questioned the positive role of partner-aware XAI in critical scenarios, such as tutoring.

In Chapter 4 we addressed our RQ1, which regarded the extent by which partner-aware explanations influence people compared to classical ones during collaborative decision-making. We showed how robots that generate explanations from their common ground with their partners can be more persuasive than robots producing more precise explanations. Interestingly, this influence was more pronounced among less skilled players, with participants not fully conscious of the robot's exploitation of interaction history; indeed, they did not express preferences for a particular kind of explanation. Future investigations could delve into the role of the robot's embodiment and social behaviour in influencing participants' decision-making.

Chapter 5 focused on RQ2, which regarded how people's personality traits contribute to HRI dynamics with explainable robots during collaborative decision-making. We observed well-known social mechanisms, such as the tendency to agree with a social partner, only with the explainable robot. Moreover, a learning effect emerged, attributable not solely to explanations but also to the participants' familiarity with the robot's playing style. Also in this regard, it will be worth investigating the role of the robot's embodiment and social behaviour in eliciting these mechanisms.

In Chapter 6, we introduced an assessment task to objectively and quantitatively measure the goodness of explanations, intended as the amount of information they provide to non-expert users. We emphasised the importance of checking whether partner-aware explanations were more useful or preferred for affective reasons. We implemented the assessment in a peculiar task and provided general criteria to design other assessment tasks that could share similar characteristics. Finally, we provided instructions on how to implement the assessment task in different HRI/HCI scenarios.

Tackling RQ3 and RQ4, Chapters 7 and 8 presented the results of the experiments involving the assessment task, comparing classical vs. partner-aware XAI approaches across different artificial agents (a virtual dialogical agent and a social humanoid robot). These research questions regarded the partner-aware explanations' informativeness for non-expert users compared to classical explanation approaches, and the influence of expert explainable agents during learning-by-doing tasks, respectively. Notably, expert artificial agents influenced participants' behaviour in task exploration during training, prompting consideration for future studies exploring the influence of expert artificial agents encouraging participant autonomy. Indeed, participants who received no assistance revealed higher exploration abilities that improved their learning.

In light of this work, it is imperative to dedicate efforts to designing explainable social robots that augment human abilities without overshadowing them. Upholding Jakob Nielsen's third usability heuristic of user control and freedom (Nielsen, 2005) is crucial. In my opinion, it is critical for the HRI community to prioritise conscious and aware interactions with social robots. One possible solution, which needs further investigation, rely on the implementation of cognitive forcing strategies in the collaborative interaction between humans and expert AI agents.

The field of explainable robotics is still in its infancy, and many challenges still need to be addressed. Key among them is the design of partner-aware explanation methods tailored for robotics, as existing XAI may not align with the unique needs of robots. Additionally, such approaches need to be tested through user studies, and it is challenging to conduct them in a controlled and rigorous way within the HRI field. Finally, I want to highlight the challenges of decoupling the XAI goodness assessment to its application context, AI model, and intended users.

This thesis has addressed some of these challenges, sparking fruitful discussions about the social dimensions of explainability, robot influence, and partner-aware XAI. Overall, the future of explainability in HRI is bright, fueled by the growing number and quality of multidisciplinary research. Further research in this area needs to continue to develop and improve XAI methods and address the limitations of the present ones. Moreover, we should prioritise a deeper understanding of the social dynamics between humans and explainable social robots and a thorough exploration of the ethical considerations in developing such robots in collaborative settings. In closing, echoing Prof. A.M. Turing, *"we can only see a short distance ahead, but we can see plenty there that needs to be done"* (Turing, 1950).

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–18, New York, NY, USA. Association for Computing Machinery.

Abdulrahman, A., Richards, D., and Bilgin, A. A. (2022). Exploring the influence of a user-specific explainable virtual advisor on health behaviour change intentions. *Autonomous Agents and Multi-Agent Systems*, 36(1):25.

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Ahmad, M., Mubin, O., and Orlando, J. (2017). A systematic review of adaptivity in human-robot interaction. *Multimodal Technologies and Interaction*, 1(3).

Amir, D. and Amir, O. (2018). Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 1168–1176, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Angelopoulos, G., Imparato, P., Rossi, A., and Rossi, S. (2024). Using theory of mind in explanations for fostering transparency in human-robot interaction. In Ali, A. A., Cabibihan, J.-J., Meskin, N., Rossi, S., Jiang, W., He, H., and Ge, S. S., editors, *Social Robotics*, pages 394–405, Singapore. Springer Nature Singapore.

Anjomshoae, S., Najjar, A., Calvaresi, D., and Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 1078–1088, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Arnold, T., Kasenberg, D., and Scheutz, M. (2021). Explaining in time: Meeting interactive standards of explanation for robotic systems. *Journal of Human-Robot Interaction*, 10(3).

Aron, A., Aron, E. N., and Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology*, 63(4):596.

Awasthi, A. and Sarawagi, S. (2019). Continual learning with neural networks: A review. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 362–365.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58.

Belgiovine, G., Gonzalez-Billandon, J., Sandini, G., Rea, F., and Sciutti, A. (2022). Towards an hri tutoring framework for long-term personalization and real-time adaptation. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22 Adjunct, page 139–145, New York, NY, USA. Association for Computing Machinery.

Bertrand, A., Belloum, R., Eagan, J. R., and Maxwell, W. (2022a). How cognitive biases affect xai-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 78–91, New York, NY, USA. Association for Computing Machinery.

Bertrand, A., Belloum, R., Eagan, J. R., and Maxwell, W. (2022b). How cognitive biases affect xai-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society*, pages 78–91.

Böckle, M., Yeboah-Antwi, K., and Kouris, I. (2021). Can you trust the black box? the effect of personality traits on trust in ai-enabled user interfaces. In *International Conference on Human-Computer Interaction*, pages 3–20. Springer.

Borgo, R., Cashmore, M., and Magazzeni, D. (2018). Towards providing explanations for ai planner decisions. *arXiv preprint arXiv:1810.06338*.

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. (2012). A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43.

Buhrman, H. and de Wolf, R. (2002). Complexity measures and decision tree complexity: a survey. *Theoretical Computer Science*, 288(1):21–43. Complexity and Logic.

Cawsey, A. (1993). User modelling in interactive explanations. *User Modeling and User-Adapted Interaction*, 3(3):221–247.

Chakraborti, T., Kulkarni, A., Sreedharan, S., Smith, D. E., and Kambhampati, S. (2019). Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In *Proceedings of the international conference on automated planning and scheduling*, volume 29, pages 86–96.

Chakraborti, T., Sreedharan, S., Zhang, Y., and Kambhampati, S. (2017). Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 156–163. AAAI Press.

Chin-Parker, S. and Bradner, A. (2010). Background shifts affect explanatory style: How a pragmatic theory of explanation accounts for background effects in the generation of explanations. *Cognitive processing*, 11(3):227–249.

Ciardo, F., De Tommaso, D., Beyer, F., and Wykowska, A. (2018). Reduced sense of agency in human-robot interaction. In Ge, S. S., Cabibihan, J.-J., Salichs, M. A., Broadbent, E., He, H., Wagner, A. R., and Castro-González, Á., editors, *Social Robotics*, pages 441–450, Cham. Springer International Publishing.

Ciatto, G., Schumacher, M. I., Omicini, A., and Calvaresi, D. (2020). Agent-based explanations in ai: Towards an abstract framework. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 3–20. Springer International Publishing.

Cohausz, L. (2022). Towards real interpretability of student success prediction combining methods of xai and social science. *Proceedings of the 15th International Conference on Educational Data Mining*, pages 361–367.

Conati, C., Barral, O., Putnam, V., and Rieger, L. (2021). Toward personalized xai: A case study in intelligent tutoring systems. *Artificial Intelligence*, 298:103503.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

De Graaf, M. M. and Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

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

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Došilović, F. K., Brčić, M., and Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215.

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

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., and Weisz, J. D. (2021). *Expanding Explainability: Towards Social Transparency in AI Systems*. Association for Computing Machinery.

Embarak, O. H. (2022). Internet of behaviour (iob)-based ai models for personalized smart education systems. *Procedia Computer Science*, 203:103–110.

Esterwood, C. and Robert, L. P. (2021). A systematic review of human and robot personality in health care human-robot interaction. *Frontiers in Robotics and AI*, 8.

Fair, D. (1979). Causation and the flow of energy. *Erkenntnis*, 14(3):219–250.

Ferrari, F., Paladino, M. P., and Jetten, J. (2016). Blurring human–machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics*, 8(2):287–302.

Ferreira, J. J. and Monteiro, M. (2021). The human-ai relationship in decision-making: Ai explanation to support people on justifying their decisions. *arXiv preprint arXiv:2102.05460*.

Fiok, K., Farahani, F. V., Karwowski, W., and Ahram, T. (2022). Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation*, 19(2):133–144.

Fisher, J. B., Robrecht, A. S., Kopp, S., and Rohlfing, K. J. (2023). Exploring the semantic dialogue patterns of explanations - a case study of game explanations. In *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue,(Marilough/SemDial'23)*.

Fiske, S. T., Cuddy, A. J., and Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83.

Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6):878.

Gächter, S., Starmer, C., and Tufano, F. (2015). Measuring the closeness of relationships: a comprehensive evaluation of the'inclusion of the other in the self'scale. *PloS one*, 10(6):e0129478.

Gambino, A. and Liu, B. (2022). Considering the context to build theory in hci, hri, and hmc: Explicating differences in processes of communication and socialization with social technologies. *Human-Machine Communication*, 4:111–130.

Gessl, A. S., Schlögl, S., and Mevenkamp, N. (2019). On the perceptions and acceptance of artificially intelligent robotics and the psychology of the future elderly. *Behaviour & Information Technology*, 38(11):1068–1087.

Ghai, B., Liao, Q. V., Zhang, Y., Bellamy, R., and Mueller, K. (2021). Explainable active learning (xal): Toward ai explanations as interfaces for machine teachers. *Proceedings ACM Human-Computer Interaction*, 4(CSCW3).

Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89.

Girotto, V., Legrenzi, P., and Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, 78(1-3):111–133.

Gong, Z. and Zhang, Y. (2018). Behavior explanation as intention signaling in human-robot teaming. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1005–1011.

Gosling, S. D., Rentfrow, P. J., and Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR.

Gray, H. M., Gray, K., and Wegner, D. M. (2007). Dimensions of mind perception. *science*, 315(5812):619–619.

Green, B. and Chen, Y. (2019a). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 90–99, New York, NY, USA. Association for Computing Machinery.

Green, B. and Chen, Y. (2019b). The principles and limits of algorithm-in-the-loop decision making. *Procedeeings of the ACM on Human-Computer Interaction*, 3(CSCW).

Grice, H. P. (1975). Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Groß, A., Singh, A., Banh, N. C., Richter, B., Scharlau, I., Rohlfing, K. J., and Wrede, B. (2023). Scaffolding the human partner by contrastive guidance in an explanatory human-robot dialogue. *Frontiers in Robotics and AI*, 10.

Gunning, D. and Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58.

Halpern, J. Y. and Pearl, J. (2005). Causes and explanations: A structural-model approach. part i: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887.

Haring, K. S., Silvera-Tawil, D., Watanabe, K., and Velonaki, M. (2016). The influence of robot appearance and interactive ability in hri: A cross-cultural study. In Agah, A., Cabibihan, J.-J., Howard, A. M., Salichs, M. A., and He, H., editors, *Social Robotics*, pages 392–401, Cham. Springer International Publishing.

Harman, G. H. (1965). The inference to the best explanation. *The philosophical review*, 74(1):88–95.

Hashemian, M., Paiva, A., Mascarenhas, S., Santos, P. A., and Prada, R. (2019). The power to persuade: a study of social power in human-robot interaction. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–8.

Heere, B., James, J., Yoshida, M., and Scremin, G. (2011). The effect of associated group identities on team identity. *Journal of Sport Management*, 25(6):606 – 621.

Heere, B. and James, J. D. (2007). Stepping outside the lines: Developing a multi-dimensional team identity scale based on social identity theory. *Sport Management Review*, 10(1):65–91.

Heerink, M., Kröse, B., Evers, V., and Wielinga, B. (2010a). Assessing acceptance of assistive social agent technology by older adults: the almere model. *International journal of social robotics*, 2(4):361–375.

Heerink, M., Kröse, B., Evers, V., and Wielinga, B. (2010b). Relating conversational expressiveness to social presence and acceptance of an assistive social robot. *Virtual reality*, 14(1):77–84.

Heider, F. (2013). *The psychology of interpersonal relations*. Psychology Press.

Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259.

Hellström, T. and Bensch, S. (2018). Understandable robots - what, why, and how. *Paladyn, Journal of Behavioral Robotics*, 9(1):110–123.

Hellström, T. and Bensch, S. (2018). Understandable robots - what, why, and how. *Paladyn, Journal of Behavioral Robotics*, 9(1):110–123.

Hesslow, G. (1988). The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality*, pages 11–32.

Hilton, D. (2017). *Social attribution and explanation*. Oxford University Press.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65.

Hilton, D. J. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, 2(4):273–308.

Hilton, D. J., McClure, J. J., and Slugoski, B. R. (2005). *The course of events: counterfactuals, causal sequences, and explanation*. Routledge.

Hilton, D. J. and Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*, 93(1):75.

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018). Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Holzinger, A., Carrington, A., and Müller, H. (2020). Measuring the quality of explanations: the system causability scale (scs). *KI-Künstliche Intelligenz*, 34(2):193–198.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312.

Hu, Y., Abe, N., Benallegue, M., Yamanobe, N., Venture, G., and Yoshida, E. (2022). Toward active physical human–robot interaction: Quantifying the human state during interactions. *IEEE Transactions on Human-Machine Systems*, 52(3):367–378.

Hume, D. et al. (2000). *An enquiry concerning human understanding: A critical edition*, volume 3. Oxford University Press.

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., and Kuk, G. (2022). Will algorithms blind people? the effect of explainable ai and decision-makers' experience on ai-supported decision-making in government. *Social Science Computer Review*, 40(2):478–493.

Kallina, E. (2020). Delegating agency? the effects of xai, personality traits, and the moral significance of the application on the reliance on autonomous systems: A user study.

Kaptein, F., Broekens, J., Hindriks, K., and Neerincx, M. (2017). 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.

Kashima, Y., McKintyre, A., and Clifford, P. (1998). The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology*, 1(3):289–313.

Kirsch, A. (2017). Explain to whom? putting the user in the center of explainable ai. In *Proceedings of the 1st International Workshop on Comprehensibility and Explanation in AI and ML*, pages 1–19.

Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190.

Kuhl, P. K. (1998). *Language, culture and intersubjectivity: The creation of shared perception*. Cambridge University Press.

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., and Doshi-Velez, F. (2019). An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*.

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., and Tan, C. (2021). Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*.

Leddo, J., Abelson, R. P., and Gross, P. H. (1984). Conjunctive explanations: When two reasons are better than one. *Journal of Personality and Social Psychology*, 47(5):933.

Lewis, D. (1987). Causal explanation. In *Philosophical Papers Volume II*. Oxford University Press.

Liao, Q. V. and Varshney, K. R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

Lichtenthäler, C. and Kirsch, A. (2016). Legibility of robot behavior: A literature review.

Lim, B. Y., Dey, A. K., and Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 2119–2128, New York, NY, USA. Association for Computing Machinery.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470.

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?". *Cognition*, 110(2):248–253.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive psychology*, 61(4):303–332.

Lombrozo, T. (2012). *Explanation and abductive inference*. Oxford University Press.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777. Curran Associates Inc.

Lyons, J. B., Nam, C. S., Jessup, S. A., Vo, T. Q., and Wynne, K. T. (2020). The role of individual differences as predictors of trust in autonomous security robots. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–5.

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press.

Malle, B. F. (2011). Attribution theories: How people make sense of behavior. *Theories in social psychology*, 23:72–95.

Martijn, M., Conati, C., and Verbert, K. (2022). "knowing me, knowing you": personalized explanations for a music recommender system. *User Modeling and User-Adapted Interaction*, 32(1):215–252.

Matarese, M., Cocchella, F., Rea, F., and Sciutti, A. (2023). Ex(plainable) machina: how social-implicit xai affects complex human-robot teaming tasks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11986–11993.

Matarese, M., Rea, F., and Sciutti, A. (2021a). A user-centred framework for explainable artificial intelligence in human-robot interaction. *arXiv preprint arXiv:2109.12912*.

Matarese, M., Rea, F., and Sciutti, A. (2022). Perception is only real when shared: A mathematical model for collaborative shared perception in human-robot interaction. *Frontiers in Robotics and AI*, page 166.

Matarese, M., Sciutti, A., Rea, F., and Rossi, S. (2021b). Toward robots' behavioral transparency of temporal difference reinforcement learning with a human teacher. *IEEE Transactions on Human-Machine Systems*, 51(6):578–589.

McGill, A. L. and Klein, J. G. (1993). Contrastive and counterfactual reasoning in causal judgment. *Journal of Personality and Social Psychology*, 64(6):897.

Menzies, P. and Price, H. (1993). Causation as a secondary quality. *The British Journal for the Philosophy of Science*, 44(2):187–203.

Millecamp, M., Htun, N. N., Conati, C., and Verbert, K. (2019). To explain or not to explain: The effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 397–407, New York, NY, USA. Association for Computing Machinery.

Millecamp, M., Htun, N. N., Conati, C., and Verbert, K. (2020). What's in a user? towards personalising transparency for music recommender interfaces. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 173–182, New York, NY, USA. Association for Computing Machinery.

Miller, D. T. and Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of personality and social psychology*, 59(6):1111.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Mohseni, S., Zarei, N., and Ragan, E. D. (2018a). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *arXiv preprint arXiv:1811.11839*.

Mohseni, S., Zarei, N., and Ragan, E. D. (2018b). A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839*, 1.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 607–617. Association for Computing Machinery.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2022). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164*.

Nayyar, M., Zoloty, Z., McFarland, C., and Wagner, A. R. (2020). Exploring the effect of explanations during robot-guided emergency evacuation. In *Social Robotics*, pages 13–22. Springer International Publishing.

Nesset, B., Robb, D. A., Lopes, J., and Hastie, H. (2021). Transparency in hri: Trust and decision making in the face of robot errors. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 Companion, page 313–317. Association for Computing Machinery.

Nielsen, J. (2005). Ten usability heuristics.

Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., and Gombolay, M. (2021a). The utility of explainable ai in ad hoc human-machine teaming. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 610–623. Curran Associates, Inc.

Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., and Gombolay, M. (2021b). The utility of explainable ai in ad hoc human-machine teaming. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 610–623. Curran Associates, Inc.

Pantecouteau, H. and Passera, B. (2017). Influence of human personality traits on trust in human-robot interactions. *Master Recherche IC2A Ingénierie de la Cognition, de la Création et des Apprentissages*, page 62.

Phillips, E., Zhao, X., Ullman, D., and Malle, B. F. (2018a). What is human-like? decomposing robots' human-like appearance using the anthropomorphic robot (abot) database. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 105–113, New York, NY, USA. Association for Computing Machinery.

Phillips, E., Zhao, X., Ullman, D., and Malle, B. F. (2018b). What is human-like?: Decomposing robots' human-like appearance using the anthropomorphic robot (abot) database. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 105–113. IEEE.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Putnam, V. and Conati, C. (2019). Exploring the need for explainable artificial intelligence (xai) in intelligent tutoring systems (its). In *Joint Proceedings of the ACM IUI 2019 Workshops,*.

Robert, L. (2018). Personality in the human robot interaction literature: A review and brief critique. In *Robert, LP (2018). Personality in the Human Robot Interaction Literature: A Review and Brief Critique, Proceedings of the 24th Americas Conference on Information Systems, Aug*, pages 16–18.

Robert, L. P., Alahmad, R., Esterwood, C., Kim, S., You, S., and Zhang, Q. (2020). A review of personality in human–robot interactions. *Foundations and Trends in Information Systems*, 4(2):107–212.

Robrecht, A. S. and Kopp, S. (2023). Snape: A sequential non-stationary decision process model for adaptive explanation generation. In *ICAART (1)*, pages 48–58.

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Häb-Umbach, R., et al. (2020). Explanation as a social practice: Toward a conceptual framework for the social design of ai systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3):717–728.

Roselli, C., Ciardo, F., and Wykowska, A. (2021). Intentions with actions: The role of intentionality attribution on the vicarious sense of agency in human–robot interaction. *Quarterly Journal of Experimental Psychology*.

Roth, A. M., Topin, N., Jamshidi, P., and Veloso, M. (2019). Conservative q-improvement: Reinforcement learning for an interpretable decision-tree policy. *arXiv preprint arXiv:1907.01180*.

Sanders, T. L., Wixon, T., Schafer, K. E., Chen, J. Y. C., and Hancock, P. A. (2014a). The influence of modality and transparency on trust in human-robot interaction. In *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pages 156–159.

Sanders, T. L., Wixon, T., Schafer, K. E., Chen, J. Y. C., and Hancock, P. A. (2014b). The influence of modality and transparency on trust in human-robot interaction. In *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pages 156–159.

Sanneman, L. and Shah, J. A. (2020). A situation awareness-based framework for design and evaluation of explainable ai. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 94–110, Cham. Springer International Publishing.

Saunderson, S. and Nejat, G. (2019). How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics*, 11:575–608.

Saunderson, S. and Nejat, G. (2022). Investigating strategies for robot persuasion in social human–robot interaction. *IEEE Transactions on Cybernetics*, 52(1):641–653.

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., and Satzger, G. (2022a). Should i follow ai-based advice? measuring appropriate reliance in human-ai decision-making. *arXiv preprint arXiv:2204.06916*.

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., and Vössing, M. (2022b). A meta-analysis on the utility of explainable artificial intelligence in human-ai decision-making. *arXiv preprint arXiv:2205.05126*.

Sciutti, A., Mara, M., Tagliasco, V., and Sandini, G. (2018). Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technology and Society Magazine*, 37(1):22–29.

Setchi, R., Dehkordi, M. B., and Khan, J. S. (2020). Explainable robotics in human-robot interactions. *Procedia Computer Science*, 176:3057–3066.

Sheridan, T. B. (2016). Human–robot interaction: status and challenges. *Human factors*, 58(4):525–532.

Shinde, P. P. and Shah, S. (2018). A review of machine learning and deep learning applications. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–6.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144.

Sokol, K. and Flach, P. (2020). One explanation does not fit all. *KI-Künstliche Intelligenz*, 34(2):235–250.

Sovrano, F. and Vitali, F. (2022). How to quantify the degree of explainability: Experiments and practical implications. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–9. IEEE.

Spaccatini, F., Pacilli, M. G., Giovannelli, I., Roccato, M., and Penone, G. (2019). Sexualized victims of stranger harassment and victim blaming: The moderating role of right-wing authoritarianism. *Sexuality & Culture*, 23(3):811–825.

Sreedharan, S., Chakraborti, T., and Kambhampati, S. (2021). Foundations of explanations as model reconciliation. *Artificial Intelligence*, 301:103558.

Stange, S., Hassan, T., Schröder, F., Konkol, J., and Kopp, S. (2022). Self-explaining social robots: An explainable behavior generation architecture for human-robot interaction. *Frontiers in Artificial Intelligence*, page 87.

Stange, S. and Kopp, S. (2023). Towards robots that meet users' need for explanation. In *HHAI 2023: Augmenting Human Intellect*, pages 361–365. IOS Press.

Stoffel, K. and Raileanu, L. E. (2001). Selecting optimal split-functions for large datasets. In *Research and Development in Intelligent Systems XVII*, pages 62–72. Springer.

Sukkerd, R., Simmons, R., and Garlan, D. (2018). Towards explainable multi-objective probabilistic planning. In *Proceedings of the 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems*, SEsCPS '18, page 19–25, New York, NY, USA. Association for Computing Machinery.

Tabachnik, B. and Fidell, S. (2007). Multivariate normality. *Using multivariate statistics*, 6:253.

Tabrez, A., Agrawal, S., and Hayes, B. (2019). 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. IEEE.

Tabrez, A. and Hayes, B. (2019). Improving human-robot interaction through explainable reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 751–753.

Tabrez, A. and Hayes, B. (2021). Mediating trust and influence in human-robot interaction via explainable ai.

Tapal, A., Oren, E., Dar, R., and Eitam, B. (2017). The sense of agency scale: A measure of consciously perceived control over one's mind, body, and the immediate environment. *Frontiers in psychology*, 8:1552.

Thellman, S. and Ziemke, T. (2021). The perceptual belief problem: Why explainability is a tough challenge in social robotics. *ACM Transactions on Human-Robot Interaction*, 10(3).

Tintarev, N. and Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4):399–439.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, LIX(236):433–460.

Van Bouwel, J. and Weber, E. (2002). Remote causes, bad explanations? *Journal for the Theory of Social Behaviour*, 32(4):437–449.

van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. (2021). Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291.

van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. (2021). Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404.

Vasilev, N., Mincheva, Z., and Nikolov, V. (2020). Decision tree extraction using trained neural network. In *SMARTGREENS*, pages 194–200.

Vered, M., Livni, T., Howe, P. D. L., Miller, T., and Sonenberg, L. (2023). The effects of explanations on automation bias. *Artificial Intelligence*, 322:103952.

Vilone, G. and Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Völkel, S. T., Schödel, R., Buschek, D., Stachl, C., Au, Q., Bischl, B., Bühner, M., and Hussmann, H. (2019). Opportunities and challenges of utilizing personality traits for personalization in hci. *Personalized Human-Computer Interaction*, 31.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31:841.

Wallkötter, S., Tulli, S., Castellano, G., Paiva, A., and Chetouani, M. (2021). Explainable embodied agents through social cues: A review. *ACM Transactions on Human-Robot Interaction*, 10(3).

Walton, D. (2004). A new dialectical theory of explanation. *Philosophical Explorations*, 7(1):71–89.

Wang, C. and Belardinelli, A. (2022). Investigating explainable human-robot interaction with augmented reality. In *5th International Workshop on Virtual, Augmented, and Mixed Reality for HRI*.

Wang, P. and Vasconcelos, N. (2020). Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8981–8990.

Wang, X. and Yin, M. (2021). Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 318–328, New York, NY, USA. Association for Computing Machinery.

Wang, X. and Yin, M. (2022). Effects of explanations in ai-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems (TiiS)*.

Weiner, J. (1980). Blah, a system which explains its reasoning. *Artificial intelligence*, 15(1-2):19–48.

Weitz, K., Zellner, A., and André, E. (2022). What do end-users really want? investigation of human-centered xai for mobile health apps. *arXiv preprint arXiv:2210.03506*.

Westberg, M., Zelvelder, A., and Najjar, A. (2019). A historical perspective on cognitive science and its influence on xai research. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 205–219. Springer.

Williams, O. (2021). Towards human-centred explainable ai: A systematic literature review. *Master's Thesis*.

Xiong, Z., Zhang, W., and Zhu, W. (2017). Learning decision trees with reinforcement learning. In *NIPS Workshop on Meta-Learning*.

Zelvelder, A. E., Westberg, M., and Främling, K. (2021). Assessing explainability in reinforcement learning. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 223–240. Springer.

Zonca, J., Folsø, A., and Sciutti, A. (2021). The role of reciprocity in human-robot social influence. *iScience*, 24(12):103424.

Appendix A

Supplementary Materials

A.1 Questionnaires for Studies in Part III and IV

A.1.1 Brief Big 5 Personality Traits

We used the same items presented in (Gosling et al., 2003).

"I see myself as ..."

Item Code	**Item Text**
1 Extraversion	Extraverted, enthusiastic
2 Agreeableness (R)	Critical, quarrelsome
3 Conscientiousness	Dependable, self-disciplined
4 Emotional Stability (R)	Anxious, easily upset
5 Openness to Experiences	Open to new experiences, complex
6 Extraversion (R)	Reserved, quiet
7 Agreeableness	Sympathetic, warm
8 Conscientiousness (R)	Disorganized, careless
9 Emotional Stability	Calm, emotionally stable
10 Openness to Experiences (R)	Conventional, uncreative

Table A.1 The items with (R) need to be reversed before analysis.

A.1.2 Sense of Agency

We used the items showed in (Tapal et al., 2017).

Item Code	**Item Text**
1 Sense of Positive Agency (SoPA)	I am in full control of what I do.
2 Sense of Negative Agency (SoNA)	I am just an instrument in the hands of somebody or something else.
3 SoNA	My actions just happen without my intention.
4 SoPA	I am the author of my actions.
5 SoNA	The consequences of my actions feel like they don't logically follow my actions.
6 SoNA	My movements are automatic – my body simply makes them.
7 SoNA	The outcome of my actions generally surprises me.
8 SoPA	Things I do are subject only to my free will.
9 SoPA	The decision of whether and when to act is within my hands.
10 SoNA	Nothing I do is actually voluntary.
11 SoNA	While I am in action, I feel like I am a remote-controlled robot.
12 SoPA	My behavior is planned by me from the very beginning to the very end.
13 SoPA	I am completely responsible for everything that results from my actions.

A.1.3 Warmth and Compentence

We used the same items showed in (Fiske et al., 2002).

"How much do you think iCub is ..."

Item Code	Item Text
1 Competence	Competent
2 Competence	Self-confident
3 Competence	Independent
4 Competence	Intelligent
5 Competence	Self-conscious
6 Warmth	With personality traits that makes it unique
7 Warmth	Warm
8 Warmth	Good-natured
9 Warmth	Sincere

A.1.4 Agency and Experience

We used the same items presented in (Gray et al., 2007).

"In my opinion, iCub is able to ..."

Item Code	Item Text
1 Experience	Feel pain
2 Experience	Feel pleasure
3 Experience	Feel fear
4 Experience	Feel joy
5 Agency	Plan its own action
6 Agency	Recognize emotions
7 Agency	Have self-control
8 Agency	Have its own morality

A.1.5 Likeability

We used a short version of the questionnaire used in Spaccatini et al. (2019).

"According to me, iCub is ..."

Item Code	**Item Text**
1	Cute
2	Pleasant
3	Nice
4 (R)	Annoying
5	Lovable

Table A.2 The items with (R) need to be reversed before analysis.

A.1.6 Anthropomorphism

We adapted the items showed in Ferrari et al. (2016).

"According to me ..."

Item Code	**Item Text**
1	iCub looks like a human.
2 (R)	iCub has the physical characteristics of a human being.
3	iCub had the look of a human being.
4	I could easily mistake iCub for a real person.
5 (R)	iCub had the appearance of a machine.
6	iCub reminds me of a child.

Table A.3 The items with (R) need to be reversed before analysis.

A.1.7 Satisfaction with the Explanations

We used the same items used in Conati et al. (2021).

"Examine your level of agreement with the following statements".

Item Code	**Item Text**
1 Usefulness	I would choose to have the explanations again in the future.
2 Usefulness	I am satisfied with the explanations.
3 Usefulness	The explanations were helpful to me.
4 Intrusiveness	The explanations distracted me from my task.
5 Intrusiveness	The explanations were confusing.
6 Intrusiveness	I found the explanations overwhelming.

A.1.8 Anxiety

We adapted our items from Robots UTAUT (Heerink et al., 2010a).

"As far as I am concerned ...".

Item Code	Item Text
1	If I should use the robot in a real context, I would be afraid to break something.
2	If I should use the robot in a real context, I would be afraid to make mistakes with it.
3	I was afraid to break the robot.
4	I was afraid to do mistakes with the robot.

A.1.9 Perceived Enjoyment

We adapted our items from Robots UTAUT (Heerink et al., 2010a).

"I think that ...".

Item Code	Item Text
1	I enjoy the robot talking to me.
2	I enjoy doing things with the robot.
3	I enjoyed playing with the robot.

A.1.10 Team Identity Scale

We used a short version adapted from Heere et al. (2011).

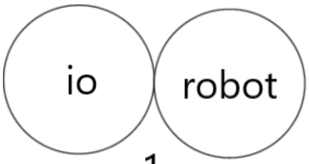
"Please indicate your degree of awareness with the following statements.".

Item Code	Item Text
1	iCub's successes are my successes.
2	I felt a strong sense of belonging to iCub.
3	Having been part of the same iCub team for me was important.

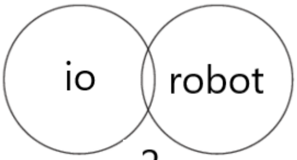
A.1.11 Inclusion of the Other in the Self (IOS)

We adapted our test from the original one from Gächter et al. (2015).

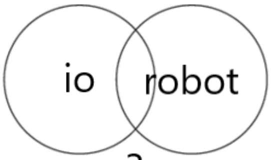
"Looking at this group of images, we now ask you to indicate which image you think best represents your relationship with the iCub robot?".



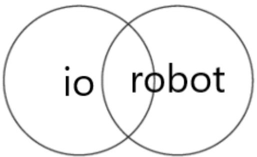
1



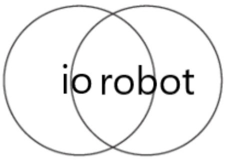
2



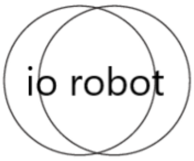
3



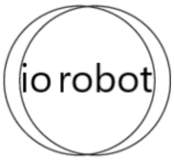
4



5



6



7