DISSERTATION

Enabling Joint Attention in Human-Robot Interaction Through Gaze

conducted in partial fulfillment of the requirements for the degree of a

Doktor der technischen Wissenschaften (Dr. techn.)

supervised by

Ao.Univ. Prof. Dipl.-Ing. Dr. techn. Markus Vincze

E376 Automation and Control Institute

co-supervised by

Assistant Prof. Dr.in phil. Mag.a phil. Astrid Weiss

E193 Institute of Visual Computing and Human-Centered Technology

submitted at the

TU Wien

Faculty of Electrical Engineering and Information Technology

by Mag.rer.nat. Michael Koller, B.Sc.

DOB 17.12.1987

Matr. Nr.: 00703506

Vienna, February 2023 Michael Koller

Acknowledgment

During my time as a Ph.D. candidate at the TU Wien I had the chance to meet many wonderful people and work on the challenging yet rewarding topic of how robots pay attention to their social surroundings.

All of this could not have been possible without the help of my supervisor, Prof. Markus Vincze, to whom I am deeply grateful for the vast amount of guidance, reflection, freedom, and motivation.

My co-supervisor, Prof. Astrid Weiss, was a tremendous help in figuring out many conceptual questions along the way. I am very thankful for receiving her mentorship and witnessing her sharp analytical thinking and heartfelt support in our many discussions.

Dr. Timothy Patten provided invaluable feedback on my work in a precise and structured manner many times, for which I am very appreciative.

I was lucky enough to be a part of two wonderful groups during my Ph.D. time, namely the Vision for Robotics research group and the Trust Robots college. A big thank you goes out to both groups for making the last years so memorable, even if the pandemic prevented us from spending more time together inside and outside work.

I also want to thank all the people I did not meet personally but who had some influence on my work. This includes the anonymous reviewers, the operational and administrative staff of TU Wien, and other organizations. This work received funding from the TrustRobots Doctoral College, EC Horizon 2020 under grant agreement No.101017089 (TraceBot) and No.665972 (ER4STEM), the Vienna Science and Technology Fund (WWTF) under project RALLI (ICT15-045), the Austrian Research Promotion Agency (FFG) under grant agreement No.879878 (K4R), and the Austrian Science Foundation (FWF) under grant agreement No.I3969-N30 (InDex).

My sincere gratitude goes out to my family, who always encouraged and supported me on my path and without whom I would not be writing these lines today. I want to say thank you to my partner, Darima, who at all times supported and believed in me throughout these years, even if I would not. I could not have done it without you.

Abstract

Humans employ gaze to coordinate their actions in joint attention scenarios. Collaborative service robots must leverage this communicative modality to fluently interact with humans during joint action tasks. They must signal their own attentional focus, thus communicating their intentions and goals, and process social cues relating to the collaborator's attentional focus and goal.

This opens up a multifaceted problem space and psychological research has revealed several distinct constituting components of joint attention. Various behavioral and cognitive aspects are categorized on different temporal resolution levels, from short-term behaviors such as gaze aversion, up to long-term cognitive capabilities such as Theory of Mind. For these phenomena, different scientific and technological research approaches are applicable. The respective findings must be integrated to arrive at a working implementation on a robotic system.

This thesis presents findings for multiple aspects of joint attention in human-robot interaction. It discusses their interrelation with respect to temporal resolution, conceptual challenges, mappings between human and technological cognitive capabilities, and how to leverage technological approaches to emulate human joint attention capabilities.

Empirical human-robot interaction research is performed to determine whether deviations from human-inspired gaze parameters are viable for robots without deteriorating the interaction with a human. A subsequent review of psychological research informs the design of a stochastic gaze controller derived from human-human interaction data during successful joint action tasks.

Through a novel algorithmic approach, plan recognition from video data in manipulation-heavy task domains is made possible while only relying on standard robotic systems such as object detection and classical planning. A virtual reality simulator and dataset provide samples of complex, long time-horizon object manipulation tasks with detailed annotation, including multiple image sequences, object poses, and logical predicates.

Our novel algorithmic approach to an expanded setting of the assistive multi-armed bandit problem improves human-robot team performance when the human acts according to an empirically documented systematic irrational bias.

We discuss the interrelation of the different contributions and propose methods for their integration. Throughout the thesis, we show how ongoing concerns about the robotic research setting, the use-case scenario cannot be disregarded. Design assumptions and interaction aspects outside of the given research setting must be critically evaluated in order to emulate the breadth and depth of human social interactions.

Kurzzusammenfassung

Menschen benutzen ihr Blickverhalten, um ihre Handlungen in *Joint Attention* Szenarien zu koordinieren. Kollaborative Serviceroboter sollen diese kommunikative Modalität ebenfalls nutzen, um mit Menschen während gemeinsamer Aufgaben flüssig zu interagieren. Sie müssen ihren eigenen Aufmerksamkeitsfokus signalisieren um ihre Absichten und Ziele zu kommunizieren. Gleichermaßen müssen Roboter auch soziale Hinweise der interagierenden Person wahrnehmen um ihren Aufmerksamkeitsfokus und ihre Ziele daraus abzuleiten.

Dies eröffnet ein vielschichtiges Problemfeld, und die psychologische Forschung hat mehrere Komponenten der *Joint Attention* beschrieben. Verschiedene verhaltensbezogene und kognitive Aspekte werden auf unterschiedlichen zeitlichen Auflösungsebenen kategorisiert, von kurzfristigen Verhaltensweisen wie Blickabwendung bis hin zu langfristigen kognitiven Fähigkeiten wie Theory of Mind. Für diese Phänomene sind unterschiedliche wissenschaftliche und technologische Forschungsansätze anwendbar. Die jeweiligen Erkenntnisse müssen integriert werden, um zu einer funktionierenden Roboterimplementierung zu gelangen.

In dieser Arbeit werden Erkenntnisse zu verschiedenen Aspekten der *Joint Attention* in der Mensch-Roboter-Interaktion vorgestellt. Es werden deren Zusammenhänge erörtert in Bezug auf die zeitliche Auflösung, konzeptionelle Herausforderungen, Abbildungen zwischenmenschlichen und technologischen kognitiven Fähigkeiten sowie die Nutzung technologischer Ansätze zur Nachahmung menschlicher *Joint Attention*-Fähigkeiten.

Wir führen empirische Forschung zur Mensch-Roboter-Interaktion durch um festzustellen, ob Abweichungen von den vom Menschen inspirierten Blickparametern für Roboter anwendbar sind, ohne die Interaktion mit dem Menschen zu beeinträchtigen. Eine anschließende Analyse der psychologischen Forschung dient als Grundlage für die Entwicklung einer stochastischen Blicksteuerung, die aus Interaktionen zwischen Menschen während gemeinsamer Aktionen abgeleitet wurde.

Durch einen neuartigen algorithmischen Ansatz wird die Planerkennung aus Videodaten in manipulationslastigen Aufgabenbereichen ermöglicht, wobei lediglich auf Standard-Robotersysteme wie Objekterkennung und klassische Planung zurückgegriffen wird. Ein *Virtual-Reality*-Simulator und ein Datensatz liefern Daten für komplexe Objektmanipulationsaufgaben mit langem Zeithorizont und detaillierter Annotation, einschließlich mehrerer Bildsequenzen, Objektposen und logischer Prädikate.

Unser neuartiger algorithmischer Ansatz für eine erweiterte Variante des assistiven mehrarmigen Banditen verbessert die Leistung des Mensch-Roboter-Teams, wenn der Mensch gemäß einer systematischen irrationalen Verzerrung aus der Literatur handelt.

Wir erörtern die Wechselbeziehung zwischen den verschiedenen Beiträgen und schlagen Methoden zu deren Integration vor. In dieser Arbeit wird aufgezeigt, dass Bedenken bezüglich des robotischen Anwendungsszenarios in wissenschaftlichen Studien und im echten Einsatz nicht außer Acht gelassen werden dürfen. Die Entwurfsannahmen und Interaktionsaspekte abseits der Forschungsfrage müssen ebenso kritisch betrachtet werden, um die Breite und Tiefe menschlicher sozialer Interaktionen zu emulieren.

Contents

Chapter 1

Introduction

> "The simple act of paying attention can take you a long way."
>
> Keanu Reeves

The field of social robotics is growing quickly [1], [2], and one of its core visions is to equip artificial physical agents with the capability of interacting naturally with humans, their surrounding, and among themselves [3]. In deliberate, intentional human behavior, attention to relevant aspects of the environment is a ubiquitous phenomenon. Psychological research has yielded several different theoretical models of attention, e.g., the spotlight model [4], the bottleneck model [5], the filter theory [6], and many more.

When two persons decide to work together on a physical task, they need to be able to coordinate their actions [7]. Coordination is necessary for both independent and joint actions. On the one hand, independent actions can be performed by a single actor, but often there must be a convention or negotiation about who performs which action. Joint actions, on the other hand, involve both actors. Throughout the whole action, they must coordinate their individual efforts. Most coordination efforts will include a third external locus of attention, e.g., the object used to perform a specific independent action or the object used throughout a joint action.

One necessary cognitive mechanism for such interactions is called joint attention [8], [9]. Successful joint attention leads to correct belief updates between the actors (e.g., negotiating who does what) and enables joint actions. Joint attention depends on the fluent interplay of many conceptual components. Actors must be able to initiate (IJA), respond to (RJA), and ensure joint attention (EJA) [10], [11]. Joint attention is embedded in a hierarchy of different levels of cognitive complexity. This hierarchy starts with the recognition of eyes, faces, gaze direction, and motion. It culminates in the cognitive ability to ascribe beliefs and intentions to other actors, formulated in the Theory of Mind (ToM) [12].

Another important aspect of joint attention is the communication modality. Joint attention is not only achieved through speech but also (both referential and mutual) gaze and gestures. The necessity to constantly coordinate joint attention in short intervals makes verbal communication relatively intrusive and cumbersome. In [13], the

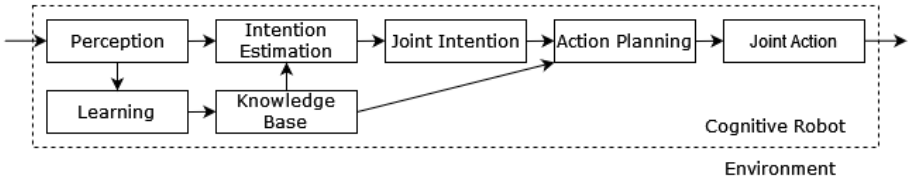

Figure 1.1: From [13]: Overview of mechanisms leading to joint action in a cognitive robot.

authors argue that fluent collaboration channels are non-verbal (e.g., gaze) and implicit (e.g., manipulation gestures that an observer interprets). Similar to structural models of attention in psychological research [14], there are structural models of joint action in human-robot interaction (HRI) [13] (Figure 1.1).

The ubiquity of attention processes in everyday human life necessitates the development of joint attention capabilities in social robots. This thesis focuses specifically on joint attention. We argue that in order to be perceived as social actors with agency, social service robots in domestic settings must be able to engage with humans by signaling their attention and to understand human attentional cues.

So far, commercially available social robots have failed to provide a sustainable long-term benefit to their owners and were shortly after purchase reduced to party tricks and expensive toys [15].[1] Arguably, one factor of the failure of these robots is the broken promise of social interaction with humans. Service robots that fulfill a practical function in a household, like cleaning the floor, struggle less with human acceptance than social robots [16].

However, as service robots fulfill more and more complex household tasks, the HRI aspect will also be more pronounced. Universal household service robots cannot act purely with a task-based rationale (e.g., a floor cleaning robot like Roomba[2]), but must be able to fluently interact with humans, e.g., in a table setting scenario. This is only possible when robots possess a functional equivalent of the human joint attention cognitive mechanism. Although there has been a vast array of fruitful work, robotic joint attention in HRI scenarios remains a milestone to be conquererd in the robotic research community.

1.1 Problem Statement

First, since joint attention, and attention processes in general, are fundamental to all conscious acting, we want to thematically distance the work presented in this thesis from artificial general intelligence (AGI) and general cognitive frameworks (e.g., SOAR [17] or ACT-R [18]) at this point. In the approaches above, the hope is that complex processes like joint attention are emergent. However, this thesis focuses on more mature approaches and their integrability on a robotic platform.

Second, we found that a useful perspective on the multifaceted problem of joint

[1] https://spectrum.ieee.org/anki-jibo-and-kuri-what-we-can-learn-from-social-robotics-failures
[2] https://www.irobot.com/en_US/roomba.html

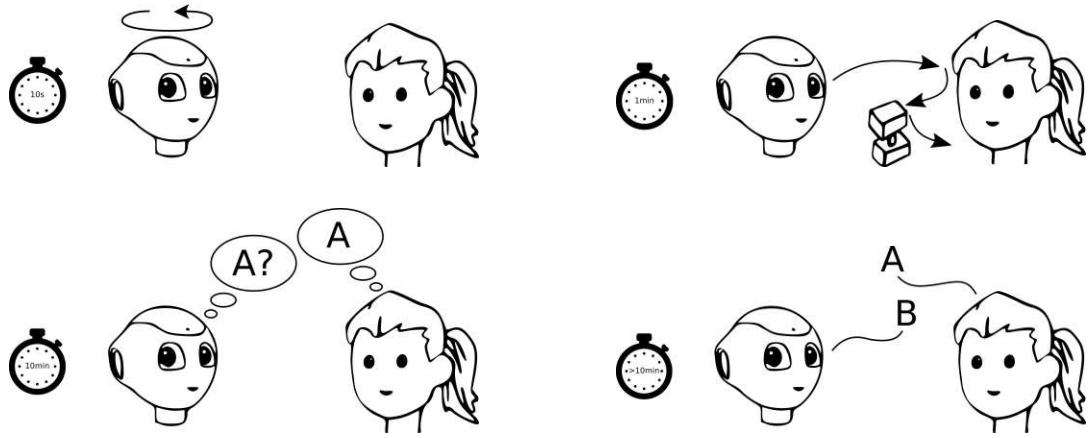

Figure 1.2: The four time resolutions of joint attention HRI scenarios in this thesis. Top left: Gaze aversion in the time resolution of up to ten seconds. Top right: Gaze sequence planning in the time resolution of up to one minute. Bottom left: Plan recognition in the time resolution of up to ten minutes. Bottom right: Preferences and biases in the time resolution ten minutes and above.

Figure 1.3: Three anthropomorphic service robots capable of object manipulation. Left: Toyota Human Support Robot (HSR). Middle: Baxter by rethink robotics. Right: PR2 by WillowGarage.

attention in social robotics is to regard the contributing mechanisms with respect to their time horizon. This abstraction allows the different aspects of attention to be broadly organized on a time resolution scale (Figure 1.2). The time resolution is high for the faster, low-level processes and concerns short time spans. In our case, this includes single atomic actions or observational events in the different communication modalities, like single gaze actions and utterances. The time resolution is low for the slower, high-level processes, which have a long duration. In our case, this includes intentions, beliefs, and preferences that take some time to form and change through repeated interactions.

Our robot scenario is a domestic service robot that interacts with humans during one of their intended tasks, such as setting the table, cleaning a space, and preparing food. This is in accordance with a commonly shared vision in the robotic research community [19], where robots are cooperating with us in our daily (personal) lives.

Many service robots possess some degree of anthropomorphism, e.g., Toyota HSR,

rethink robotics Baxter, or Willow Garage PR2, through the fact that most optical sensors are mounted in a head-like camera mount on top of the robot torso (Figure 1.3). In line with [20], we argue that such robots should have the same limitations as a human if the embodiment is humanoid in order not to betray the trust of the human collaborator. E.g., if a humanoid robot moves its head and eyes in a specific direction, the implication for lay people is that the robot pays attention to something in the general direction of the gaze. However, a robot could have more cameras mounted in its torso or back of the head to monitor its whole surrounding. This would betray the expectations of the human collaborator if they are not yet familiar with such a robot.

Future domestic service robots are intended for long-term deployment, which leads to additional challenges, such as identifying the start and end of relevant tasks that humans perform in the surroundings of the robot in order to offer help for some tasks conveniently. Another opportunity in long-term deployment is the estimation of the preferences of the co-inhabiting humans in order to have helpful priors when determining how the robot will offer help. In contrast to industrial settings, households are a more dynamic, unstructured environment where robots must be able to adapt to new task instances, settings, and people. We believe that the capability of robots to follow and respect social norms during interactions will improve their efficiency and acceptance co-inhabiting humans.

1.2 Research Questions

One of the visions in the field of robotics is the development of a service robot that interacts naturally and efficiently with humans during a task in an unstructured social environment [19]. One essential ability of such a robot is to react and coordinate with other agents, such as humans, through a joint attention mechanism. In the attempt to realize this vision, as argued above, there are many possible approaches. In this thesis, we examine specifically which challenges arise when gaze-related joint attention is examined at different levels of temporal resolution. Therefore, the following research questions arise for this thesis:

- **RQ 1**: Which HRI challenges can be identified by examining gaze-related joint attention through the lens of different temporal resolutions?
- **RQ 2**: What are the relevant robotic capabilities for gaze-related joint attention in the context of different temporal resolutions in comparison to humans?
- **RQ 3**: Which existing technologies and approaches can be extended, modified, and used to improve gaze-related joint attention in different temporal resolutions?

We pose additional research questions in Chapter 2 about gaze behavior (RQ 4 a-c), other interaction behavior (RQ 5 a, b), and attitudes towards a robot (RQ 6 a-c) in conversational settings, and in Chapter 6 about whether a human-robot team can improve the performance of a biased human in scenarios when risky behavior is justified (RQ 7) and when risk-averse behavior is justified (RQ 8). These research questions are specific to the chapter and answered therein.

1.3 Contributions and Outline

We present in the following chapters our contributions of joint attention-enabling mechanisms in different time resolutions in HRI settings. The highest time resolution relevant for HRI in the context of joint attention in this thesis lies in the range of seconds. This includes single utterances, glances, and atomic manipulation actions, among others. The lowest time resolution for joint attention in HRI settings is as long as hours, days, or more generally a time frame in which beliefs about another actor can change throughout repeated interactions. The first and second time resolutions are meant to depict parts of an interaction, i.e., the specified time frame occurs multiple times during a task. The third time resolution is meant to represent the whole task duration. The last time resolution regards repeated interactions. Our contributions are thus grouped into four discrete time resolutions (Figure 1.2)

- up to 10 seconds: single gaze actions,
- up to 60 seconds: short sequences of gaze actions,
- up to 10 minutes: recognition of intentions and goals,
- 10 minutes and above: identification of preferences and biases.

This grouping serves as a structure for this thesis and is chosen with regard to the duration of different and diverse cognitive and physical actions and processes during a joint attention process. The fast processes in the high time resolution include phenomenons such as gaze following, visual object search, whereas slow processes include the formation and change of beliefs and attitudes.

The key point of this ordering is that the level of abstraction in joint attention scenarios between these time resolutions varies. Thus, the individual research questions, methods, and insights of the following chapters will apply to different time frames.

Krämer et al. [21] argue that the width and depth of human coordination capabilities in social contexts will be out of reach for the foreseeable future for technological systems and propose to restrict problem domains, simplify problems, and manage user expectations, among other strategies. We apply this recommendation by contextualizing different research problems in the relevant time resolution.

1.3.1 Gaze Aversion

In the time resolution of one to ten seconds (Figure 1.2, top left), single glances and head movements are suitable units of study. Similar to human-human interaction, gaze is an essential modality in conversational human-robot interaction settings [22]. Gaze aversion serves social purposes, e.g., intimacy regulation and turn yielding. Thus, one of the resulting descriptive parameters of an interaction is the so-called gaze aversion ratio. It describes the average amount of time spent averting the gaze from the interaction partner during an interaction.

Previously, human-inspired gaze parameters have been used to implement gaze behavior for humanoid robots in conversational settings and improve the user experience [23].

Other robotic gaze implementations disregard social aspects of gaze behavior and pursue a technical goal, e.g., face tracking [24]. However, it is unclear how deviating from human-inspired gaze parameters affects the user experience.

In Chapter 2, we report the results of an empirical HRI study, where eye-tracking, interaction duration, and self-reported attitudinal measures were used to study the impact of non-human inspired gaze timings on the user experience of the participants in a conversational setting. We show the results for systematically varying the gaze aversion ratio of a humanoid robot over a broad parameter range from almost always gazing at the human conversation partner to almost always averting the gaze.

1.3.2 Gaze Sequences

In the time resolution of about one minute (Figure 1.2, top right), the sequence of gaze targets in a joint action HRI scenario is a complex phenomenon and contains information about the actors' beliefs and intentions [25]. Gaze has more than one function in such a context. First, a humanoid robot uses the gaze to collect sensory data concerning the state of the task at hand, e.g., monitoring the state of blocks on a table during a block stacking task or the tabletop when setting the table. Second, a humanoid robot can use the gaze for social interaction, including mutual and referential gaze at objects.

We imagine that service robots must collaborate with humans in physical object manipulation tasks to assist in everyday scenarios. This collaboration requires joint attention to smoothly accomplish a shared goal smoothly. In Chapter 3, we discuss the human gaze in physical tasks and its underlying cognitive mechanisms. We present a novel probabilistic robotic gaze controller in object-centered collaborative physical tasks that allows a robot to signal its current belief state through gaze behavior for a given goal and its inclusion in a well-known joint action human-robot interaction benchmark.

1.3.3 Intentions and Goals - Plan Recognition

In the time resolution of several minutes (Figure 1.2, bottom left), complex tasks can be performed. In our case, this includes assembly tasks, or domestic tasks, such as food preparation. In this time horizon, the goals and intentions of actors are adequate units of study. In Chapter 3, we establish how a robot signals its current belief state through gaze behavior, but only if it has decided on the current goal for collaboration. In Chapter 4, we apply the same limitations, namely that no explicit action recognition, but only object detection is available, and we derive an algorithm that allows an observer to estimate an actor's current goal in a given task domain and goal set.

Plan recognition as planning is an approach that uses domain knowledge and domain-independent solvers to estimate the most likely goal in a goal set given an observation trace of atomic actions [26]. This trace is typically generated by an activity recognition procedure to determine the atomic actions within the observation trace. We propose a method that uses no explicit activity recognition but instead instantiates actions by comparing properties of interacting objects to static preconditions in planning operators.

1.3.4 Intentions and Goals - VR Dataset

In the previous chapter, [27] provided annotated samples for our plan recognition algorithm. However, their detailed annotation process was only applied to a few samples of a larger existing dataset since detailed human annotation is highly time-consuming.

To alleviate this problem, we present the *Virtual Annotated Cooking Environment* (VACE), a new open-source virtual reality dataset (`https://sites.google.com/view/vacedataset`, doi:10.48436/r5d7q-bdn48, doi:10.48436/9y2x1-q4n71) and simulator (`https://github.com/michaelkoller/vacesimulator`) for object interaction tasks in a rich kitchen environment in Chapter 5. We use the Unity-based VR simulator to create thoroughly annotated video sequences of a virtual human avatar performing food preparation activities.

1.3.5 Preferences and Biases

In the previous two chapters, our contributions aim at improving the capability of estimating an observed actor's current goal in a given domain. However, plan recognition over a given goal set incorporates a prior distribution over the goal set [26]. This distribution can be uniform if no information is available or shaped to depict the base preferences of an actor more accurately, e.g., learned through repeated prior observation of an actor. These repeated interactions are best understood in the time resolution of ten minutes and above, possibly days and months. In this time resolution (Figure 1.2, bottom right), multiple completed tasks are the adequate unit of study with a focus on the goals which were chosen within the task. As an example of such repeated tasks, one can imagine that the robot observes a human multiple times doing a specific task and noting the task outcome, e.g., which dish was prepared in the kitchen or how the table was specifically set up.

In Chapter 6, we study the assistive multi-armed bandit problem [28]. It formalizes an autonomous system that observes and intercepts the repeated actions of a human, estimates the true utility of the different actions, and potentially chooses a different action than the human to improve the overall return. This setting models team situations between a human and an autonomous system like a domestic service robot. Previous work deals with human policies in human-robot teams that are (noisily) rational or in some way communicative about the rewards. However, empirically shown human biases such as the risk-aversion described in the Cumulative Prospect Theory [29] shifts the perceived action utilities so that previous methods will only learn to repeat the bias. In this chapter, we expand the assistive multi-armed bandit setting, and derive an algorithmic approach to improve the team performance for challenging human policies.

1.4 List of Publications

The work of this thesis has previously been published or is under review in peer reviewed journals, books, conferences, and workshops.

- **M. Koller**, "Systematic variation of gaze timings and effects on the human level of comfort and feeling of being attended," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction*, IEEE, 2019, pp. 721–723.
- **M. Koller**, D. Bauer, J. de Pagter, G. Papagni, and M. Vincze, "A pilot study on determining the relation between gaze aversion and interaction experience," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction*, IEEE, 2019, pp. 644–645.
- **M. Koller**, T. Patten, and M. Vincze, "Plan Recognition from Object Detection Traces." In *AAAI Workshop on Plan, Activity, and Intent Recognition (PAIR)*, Url: http://www.planrec.org/PAIR/PAIR%2020/Resource_files/PAIR20papers.zip, 2020.
- **M. Koller**, T. Patten, and M. Vincze, "Risk-averse biased human policies in assistive multi-armed bandit settings," *TRAITS Workshop, 16th ACM/IEEE International Conference on Human-Robot Interaction*, arXiv preprint arXiv:2104.05334, 2021.
- **M. Koller**, T. Patten, and M. Vincze, "Risk-averse biased human policies with a robot assistant in multi-armed bandit settings," in *The 14th PErvasive Technologies Related to Assistive Environments Conference*, 2021, pp. 483–488.
- **M. Koller**, T. Patten, and M. Vincze, "A new vr kitchen environment for recording well annotated object interaction tasks," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction*, IEEE, 2022, pp. 629–633.
- **M. Koller**, A. Weiss, and M. Vincze, "I see what you did there: Towards a gaze mechanism for joint actions in human-robot interaction," in *Trust in Robots*, M. Vincze and S. Koeszegi, Eds., Vienna: TU Wien Academic Press, 2023, pp. 149–178.
- **M. Koller**, A. Weiss, M. Hirschmanner, and M. Vincze, "Robotic gaze and human views - a systematic exploration of robotic gaze aversion and its effects on human behaviors and attitudes," *Frontiers in Robotics and AI*, vol. 10, 2023.

This thesis is structured in the following way. The next five chapters describe our contributions in the order presented in the outline above. In the discussion, we answer our overarching research questions and discuss and integrate the findings in the context of the thesis. We conclude this thesis with a summary, limitations, and future work.

Chapter 2

Gaze Aversion

In HRI settings, such as visual joint attention tasks or human-robot conversations, both robotic and human gaze guide the attention of the interaction partner, influence the attitude towards them, and provide and shape the rhythm of the interaction. Therefore, how engineers integrate robotic gaze into HRI settings is an impactful design decision.

Successful joint attention depends on all involved actors' capability to signal their own and detect the other's locus of attention. Signaling and detecting occur in a loop throughout the interaction, and one iteration of this loop usually lasts a few seconds [30] (Figure 2.1). Thus, much time in a joint attention task is spent gazing at the other actor to determine their current locus of attention, as well as designing one's gaze behavior in a way that is readable to the other. Humans exceed in this task to such a degree that an observer can gain relevant information even from the natural task-related gaze of the other [31].

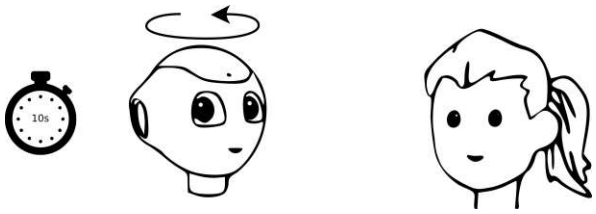

Figure 2.1: In the conversational HRI setting, in the time resolution of up to ten seconds, robotic gaze aversion has an effect on the human interaction partner.

Additionally, humans actively modify their gaze behavior as a signal to a collaborator, so it carries enough task-related information, e.g., referentially gazing at the next task-related object until the other actor realizes this is an important object. The robotic task-related gaze in current HRI scenarios is often designed to achieve the given task goals. In joint action tabletop scenarios involving a humanoid service robot and a human, the robot's gaze is usually statically directed at the task space, i.e., the table with all relevant objects [32]. This robot behavior maximizes the image processing capabilities of the tested system but neglects the gaze-related joint attention communication between the two actors. Neither does the robot react to the human

gaze nor does it signal any intentions through its gaze. [1]

On the one hand, using the robot gaze to maximize the image processing of the joint action task and, on the other hand, using the robot gaze to signal and detect joint attention are conflicting goals. They can be reconciled by improving either aspect on a technological basis or finding the right balance between the time spent gazing at the task space and the other actor. However, if these processes are sufficiently fast, the underlying question of how to split the gaze between targets for an optimal human interaction experience will remain relevant.

Furthermore, the right gaze timing that feels comfortable to the human action partner and the optimal robot gaze timings could differ from usual human-inspired gaze timings [22]. Also, human-inspired gaze timings can not be transferred precisely to humanoid robots since their embodiments have different constraints than human bodies. For example, there might be too few degrees of freedom and motor restrictions. However, even if currently available robots might not succeed in imitating human movement, this might also not be necessary in the first place since humans can also interpret and feel comfortable with non-human-inspired motions [34]. The optimal gaze parameter settings for a humanoid robot could be close to human parameters, but not necessarily.

Previous work has yet to examine this balance between social signaling and task-specific gaze. Robot designers will profit from knowing how much freedom they have in the design of an interaction when they need to balance the face-directed or task-directed gaze. To focus on the social aspect of that balance, we reformulate the research questions in the context of social conversational settings, i.e., settings where there is no joint action task. In such a setting, gazing away from the interaction partner has different meanings, e.g., inattentiveness or focus on something in the surroundings. Even for this setting, it still needs to be determined whether human-inspired gaze timings are the best choice for humanoid robots.

Previous work often compares one well-designed gaze behavior against one intentionally poorly designed one or a neutral one. This method is valid to arrive at a high-performing gaze behavior that results in high human comfort and other trust-related measures [22, Section 5] (e.g., [23]).

However, in these previous studies, the high-performing gaze behavior is usually modeled as closely as possible to the previously determined human gaze parameters. This means that there is a knowledge gap: How far can the gaze behavior of a humanoid robot diverge from human-inspired gaze parameters and still have an acceptable performance? This question must be posed for all gaze parameters. In the context of the second-to-second joint attention interactions, the gaze aversion ratio is an important parameter.

Gaze behavior is specified by low-level parameters, such as animation curves of head and eye movements, specific gaze targets or directions, and frequency and duration of fixations. Similarly, gaze is dependent on the current context, such as the interaction partner [35]. Previous works report a relationship between one high-level gaze behavior

[1] A common way of executing robotic grasping is to apply the following steps sequentially: Process sensor data, plan grasp trajectory, then perform grasp, and finally check if the grasp attempt was successful. Sensor input, in this case, is often only used in the first processing step. This operational schema would allow the robotic head to be used for signaling purposes when the sensor input is not needed. However, this approach is neither very robust (cfg. active vision [33]) nor human-like.

parameter, namely the *gaze aversion ratio* (GAR), and the perceived *user experience* (UX) [36]. A positive UX is central for accepting social robots into our everyday lives [37]. In current studies with social humanoid robots such as Pepper[2], the GAR for a positive UX is often determined through a human-inspired (e.g., [23]) approach: Gaze timings in Human-Human Interaction (HHI) are recorded and replicated as best as possible in robotic systems.

We consider GAR a central parameter in conversational HRI settings since it is a good representation of the overall gaze behavior of a robot. Even if the robot behavior designer does not explicitly model the behavior around a chosen GAR, the GAR of a behavior can always be computed. Gaze aversion is also a social interaction parameter for which a broad body of HRI work exists [22]. A common approach is to create a human-inspired gaze behavior. Thus, deviating from such human-inspired gaze parameters produces relevant complementary findings to the human-inspired approach.

Additionally, deviating from human gaze behavior is necessary if the robotic gaze has specific additional goals and limitations. This includes movement restrictions imposed by the robot embodiment or sensor limitations that impose minimum observation times for detecting certain objects [24] if the robot's gaze coincides with the camera view. This problem is relevant if design principles of *honest anthropomorphism* are followed, e.g., that the camera is installed in the head of the robot in its assumed gaze direction [20]. This ethical consideration is reinforced by practical limitations of purely animated robotic eye movement. In [38] the authors report that humans are less accurate in interpreting the gaze direction through eye movement of animated robotic faces in contrast to real human faces.

The effects of deviating from such HHI-inspired GAR have not yet been systematically studied, and there are unexplored assumptions: (1) Is the human-inspired parameter setting optimal for a specific robot embodiment and scenario? Other parameter settings might be equally or more appropriate for a specific robot embodiment, and these will never be identified when adhering to human-inspired parameters as closely as possible. (2) Assuming the HHI parameter setting is optimal, how much does a deviation from that optimum degrade the UX? Insights to these questions benefit robot designers who need to balance social and technical aspects of robotic gaze behavior.

In this chapter, we explore a broad range of values for the robotic GAR for the Pepper robot and whether there are systematic effects on the different dependent interaction measures. We chose this parameter because it is a high-level gaze behavior descriptor and can be derived for every robotic social gaze behavior. Thus, we argue that the findings of this chapter are relevant to many different experimental settings. There are different implications for future robot-centered and human-centered research, depending on the outcome: If there is indeed only one high-performing setting that can be interpreted as "close to the human behavior", it might be reasonable to only inspect settings close to this behavior in the future. Otherwise, if we observe several high-performing settings or negligible differences, robot designers will have more freedom in their choices. For technology-focused approaches, if there is knowledge about more lenient acceptable gaze timings, it is easier for robot designers to approach the range of

[2]https://www.softbankrobotics.com/emea/en/pepper

acceptable parameters in their implementation.

We created an experimental design to determine the GAR that enables a positive UX without being derived from HHI. As measures for UX, we chose *feeling of being attended*, *feeling of comfort within the interaction* and *perceived interaction capabilities of the robot* [39]. Furthermore, we recorded the gaze behavior of the participants as an implicit measure of UX since human gaze behavior (including gaze aversion [40]) is closely linked to affect and emotion. We additionally measure other interaction parameters, namely the *interaction duration* and *word count* uttered by the human participants, as these variables have been linked to affective, emotional, and attentional states [41]. We found that the robot gaze behavior shapes human behavior. Broadly, humans mimicked the gaze aversion ratio of the robot but did not do so through mere gaze following. However, we did not find that different gaze aversion parameter settings resulted in significant differences in self-reported attitudes toward the robot. This means that humans - at least consciously - did not find any interaction scenarios more or less comfortable than the other. This result allows robot designers to consider the balance between task-related and social gaze for joint attention more freely than before in the second-to-second time scale of a social interaction.

In this chapter we present the following contributions:

1. We introduce an experimental design to vary one parameter of robot behavior, namely the GAR, in a wide range of possible values for a specific anthropomorphic robot, namely Pepper, without assuming that human behavior is optimal.

2. We implement a minimal-animacy robot behavior that isolates the effect of the mere ratio of gazing at the human conversation partner vs. averting the gaze to the side.

3. We conducted a user study ($n = 101$) in which we recorded, evaluated, and interpreted a rich dataset (doi:10.48436/frswc-4dn44) composed of gaze-tracking records over the whole duration of the interaction, other behavioral measures, such as spoken word count and interaction duration, and self-reported attitudes toward a social robot.

4. We discuss guidelines for implementing robot behavior that adheres less strictly to human-derived parameters.

The remainder of this chapter is organized as follows: Related work about gaze and eye-tracking in previous HRI experiments is presented in Chapter 2.1. In Chapter 2.2, the research questions are posed, followed by the used methods in Chapter 2.3. Next, results are presented in Chapter 2.4 and discussed and transferred into guidelines in Chapter 2.5. We finish with a conclusion in Chapter 2.6. This chapter's content is based on previously published work in [42]–[44].

2.1 Related Work

In this chapter, we first present related work in both human and robotic gaze behavior research. We review different forms of social gaze and why it is a crucial nonverbal social

modality for humans. Then we review how social gaze behavior has been implemented in previous HRI studies with different experimental designs. Next, we review relevant user studies incorporating eye-tracking devices and different ways of extracting empirical findings from eye gaze data. Human eye gaze reveals latent variables such as emotional and cognitive attributions towards the robot. These measures have been used in HHI, but also in HRI studies.

2.1.1 Robotic and Human Gaze Behavior in HRI

The GAR is a high-level descriptor of gaze, however, gaze aversion can occur in different circumstances with varying goals: In [45], gaze as a social cue is discussed as an evolutionary phenomenon that allows humans to interpret complex social scenarios in the following taxonomy: (1) *Mutual gaze* describes a setting, where two interaction partners are looking at each others' faces. (2) *Averted gaze* occurs when one partner looks away from the other. (3) *Gaze following* occurs when one partner notices the averted gaze of the other and then follows their line of sight to a point in space. (4) *Joint attention* is similar to gaze following, except that the person averting the gaze has a specific object as gaze target, and the other person also attends to that object. (5) *Shared attention* is a bidirectional process, where both partners simultaneously perform a mix of mutual gaze and joint attention on the same object. Thus, both know that the other is looking at the target object. (6) *Theroy of Mind* (ToM) uses higher-order cognitive strategies and performs a mental state attribution to the interaction partner. Thus one partner can determine that the other intends to interact with an object or react to a stimulus because they intend to achieve a goal by doing so or have a particular belief about it that leads to a specific action. This classification does not separate gaze aversion into a separate category. Gaze aversion can occur in the averted gaze, but also gaze following, joint attention, and shared attention. However, the GAR is still an objective, condensed descriptor of gaze behavior.

Like above, the GAR can also differ in the context of its social function and occur through different gaze acts: Five functions of social gaze and six types of gaze acts have been identified [46]. The five functions are *establishing social agency* (reinforce presence and aliveness), *communicating social attention* (show interest in human), *regulating the interaction process* (manage participation and turn taking) , *manifesting interaction content* (looking at object of interest), and *projecting mental state* (express emotions and intentions). The six gaze acts *fixation* (at target object or person), *short glance*, *gaze aversion* (away from a person), *concurrence* (repetitive horizontal or vertical movement to interrupt fixations), *confusion* (signified by consecutive rapid gaze shifts back and forth), and *scan* (several short glances to random points in space) have been identified.

Specifically for conversational settings, gaze aversion has additional functions [23]: floor management, intimacy regulation, and indication of cognitive effort. *Floor management* gaze behavior consists of gaze aversions during speech pauses to indicate that the conversational floor is being held and an interruption of the speaker is not desired. *Intimacy regulation* between two conversation partners is achieved by different degrees of gaze aversion, depending on the relationship of the conversation partners, the conversation topic, and scenario properties like the physical distance of the two

speakers. *Cognitive effort* can lead to more gaze aversions of the speaker, as they can better focus on the planning and delivery of the following utterances.

We summarize that gaze aversion behavior is highly dependent on the conversation setting, the two interaction partners, and dynamic aspects during the flow of conversation. However, there have been attempts to derive empirical average percentages for gaze at the interaction partner and mutual gaze: [47], [48] report that one person in a conversational dyad spends about 60% of the time looking at the conversation partner. 30% of the interaction time is spent in mutual gaze. While listening, people gaze more often at the conversation partner (71%) than while speaking (41%). The authors report high interpersonal variance for gaze aversion. In our experimental setting, the robot takes on the role of the listener in a dyadic conversational setting. A GAR value of 0.3 (corresponding to gazing at the partner about 70% of the time) thus constitutes the HHI standard.

Rightly, previous work focused on determining robotic gaze behavior that improves the interaction [22, Section 5]. However, the complexity of the implementation and other test-theoretic design demands would have led to an infeasibly large sample size for more fine-grained experimental setups. Thus, in the different test conditions, the presumed high-performing implementation is often compared against a deliberately poorly performing or neutral condition. This type of research question is valid for verifying specific implementations. Our work aims to answer the complementary question of which insights can be generated for future robot behavior designs when there is no explicit split between different performance groups. This is also motivated by [22, p. 37]: "It is tempting to assume that perfectly matching robot gaze behaviors to human gaze behaviors will elicit identical responses from people, but this is not always the case," as was shown in [49].

2.1.2 Eyetracking in HRI

Eyetracking in HRI research produces detailed and rich data samples during interactions and provides task-specific sensory data to robotic learning algorithms. Given the inaccuracy of other gaze estimation methods such as head pose estimation [50], including eye tracking as dependent variable into experimental designs studying conversational HRI settings is a common technique. This practice has yielded several insights into the connection between gaze behavior and UX:

[51] operationalize eye-tracking measures for the feelings and attitudes of participants over time towards different robot embodiments in a conversational HRI scenario. They found that human gaze aversion in a social chat is an indicator for the uncanniness of a robot. Similarly, in a joint task, the more a participant gazed at the robot, the worse they performed. Liking of the robot and mutual gaze develops congruently over the course of multiple interactions. Specifically, the reported uncanniness decreased, while mutual gaze by the participants increased. Also, the authors argue that later interactions represent a more stable gaze pattern after the novelty effect has worn out.

Additionally to pure conversational settings, mutual gaze has also been used to estimate social engagement of participants in joint task settings [52], [53].

[54] interpret the freely chosen interaction duration of their participants with a social

robot as an implicit measure for interest in engagement with the robot. The amount of looking at the robot is a measure for attention towards the robot. They found that in both experimental conditions (talking/gesturing robot) of a conversational object reference game, participants spent about 70% of the time looking at the robot in both conditions with no differences in the self-reported likability of the robot.

[55] employ a data collection procedure including video recording and eye-tracking of human participants in an HRI scenario, where participants taught object names to either a robot or another confederate participant. They report differences in the amount of time spent gazing at the face of another human or a robot interaction partner. Intimacy regulation [56] in a conversation is achieved by averting the gaze from the other person. Inanimate characters such as persons on TV, and social robots who exhibit limited sociability, receive more gaze than physically present human interaction partners. This result is an indication against the human-inspired gaze design approach.

[57] interpret the direction and timing of a human's gaze over time towards a robot while interacting by comparing different gaze metrics derived from temporally split interaction thirds. They notice a decrease in gaze at the robot's head over time and thus propose this measure as a proxy for engagement in the interaction and the perceived social agency of the robot.

In [58], human eye gaze acts as an implicit measure of engagement to determine whether an anthropomorphic conversational robot is gazed upon as if it was a communicative agent or a technological tool. The robot performed the task of reading a newspaper article out loud. Participants exhibited a high level of gaze toward the face of the robot. This is regarded as evidence for the communicative agent hypothesis, confirming the media equation hypothesis [59]. There was a dynamic change in gaze targets: At the beginning of an interaction, the participants exhibited more gaze behavior toward the head of the robot than at the end, although the small sample size limits the validity of this observation.

Eye gaze behavior has also been linked to persistent personality traits [60]. Analyses of HHI and HRI conversations revealed that participants gaze more at the partner's body in HRI than in HHI, more at the partner's face in HHI than in HRI, and a positive correlation between the character trait *openness* and the average duration of gazing toward partner's body in HRI. This interpersonal effect is modulated by the situational effect of *intimacy regulation* during a conversation.

[61] report a temporal effect in their conversational HRI setting, namely that people who spend more time with the robot have less favorable attitudes towards it. Participants spent, on average, 50% of the time looking at the face of the robot, with large interpersonal differences. They confirmed previous findings that the maximal pupil diameter correlates with cognitive load, and concluded that telling a story constitutes a measurable cognitive load.

[36] report that a prolonged stare of a robot at the participants increased their arousal. In their interactive gaze experimental condition, a prolonged mutual gaze of the participant at the robot correlated positively with a higher rating in fluency, fun, and connectedness.

Similarly, [62] found in HHI conversations that gaze aversions and mutual gaze are highly dependent on both interaction partners. They report a positive correlation

between mutual gaze and the combined person's agreeableness, as well as their familiarity. Concerning gaze dynamics between the two interaction partners, they report correlations implying that mutual gaze cannot be increased by only one of the two participants by simply looking at their opposite for a longer amount of time.

In summary, eye-tracking has several favorable properties for HRI research: It produces more objective data than self-report measures. The resulting data is richer and more granular than questionnaire results. Eye-tracking data reveals behavior, attitudes, and emotional states that escape the conscious verbalization of participants when filling out questionnaires. Previous studies allow us to link gaze behavior with emotional states and attitudes.

2.2 Research Qestions

As our work aimed to determine whether viable parameter ranges for HRI behavior implementations can be found without adhering to previously determined HHI parameters, we varied one gaze behavior factor, namely the Gaze Aversion Ratio (GAR). It is defined as the ratio of time averting the gaze from the interaction partner to the gaze cycle time. A GAR of 1.0 relates to always averting the gaze from the interaction partner, whereas a GAR of 0.0 relates to always gazing at the interaction partner.

Different robot behaviors can implement the same GAR since different movement profiles and gaze cycle lengths can average the same amount of time gazing at and away from the interaction partner. In our study, we wanted to establish whether the mere ratio of time spent gazing at the interaction partner's face and the time spent averting the gaze is already a significant factor for how the users experience the behavior of the robotfg, without varying different dynamics such as animation curves.

The Softbank Pepper[3] robot was chosen as the robot embodiment. The humanoid shape allows a natural implementation of gaze aversion, while the unactuated face of the robot (except for LED lit eyes) does not signal a false affordance to participants, i.e., participants do not expect elaborate facial animation during the interaction. Pepper robots are currently a widespread social robotics research platform, and thus findings can be incorporated in future research for the same platform (and similar platforms like the Softbank NAO robot [4]). Lastly, these robot platforms adhere to the *honest anthropomorphism* design principle, i.e., the robot only gathers visual information in a visual field where the anthropomorphic head is pointed. Designing robot behavior for robot embodiments that do not violate the users' trust poses additional challenges but is necessary for a trustworthy relationship.

Our main research interest in this chapter was to determine to what degree the GAR influences the behavior and attitudes of the human interaction partners. To operationalize this we derived research questions that can be grouped into three categories: 1) gaze behavior of participants; 2) interaction behavior; and 3) attitudes.

[3]https://www.softbankrobotics.com/emea/en/pepper
[4]https://www.softbankrobotics.com/emea/en/nao

1) Gaze behavior:

- **RQ 4a** Does the GAR have an effect on the fixation durations of the participants?
- **RQ 4b** Does the effect of the GAR on the fixation durations change during the interaction?
- **RQ 4c** Does the GAR influence the fixation sequence behavior of the participants?

2) Interaction behavior:

- **RQ 5a** Does the GAR have an effect on the amount of words spoken by the participants?
- **RQ 5b** Does the GAR have an effect on the interaction duration?

3) Attitudes:

- **RQ 6a** Does the GAR have an effect on the participants' perception of the robot's attention towards them?
- **RQ 6b** Does the GAR have an effect on the participants' own feeling of comfort during the interaction with the robot?
- **RQ 6c** Does the GAR have an effect on the participants' perception of the robot's interaction capabilities?

2.3 Methods

We designed an experiment in a conversational interaction setting and validated it in a pilot study [43]. During the interaction, the robot greets a human participant and asks about their favorite movie. Then it listens to their statement. When the human stops talking, the robot thanks the participant and says goodbye to conclude the interaction. In the categorization scheme of [63], the experiment was a lab-based, Wizard-of-Oz style study where a real robot and convenience-sampled participants interacted a single time.

The gaze aversion of the robot was the main behavior of the robot while the participant was talking. It was manipulated as the independent variable. The gaze aversion behavior cycled every 10 s. During a cycle, the robot acted according to a specific gaze aversion ratio (GAR) and a predetermined gaze dynamic. Figure 2.2 shows the operationalization of the independent variable GAR across five conditions (0.1, 0.3, 0.5, 0.7, 0.9).

We added additional animacy behavior since gaze aversion head shifts occur only every few seconds to avoid the robot to appear as "turned off" during the listening phase of an experiment trial. [64] report twelve animation principles which were derived from animation practices found in animated movies. These principles have also found use in the animation of robots [65] to improve aspects such as readability or likeability. There are several typically used animation profiles in HRI experiments to convey animacy. [66] implemented the following idle movements on a Baxter robot in an HRI scenario:

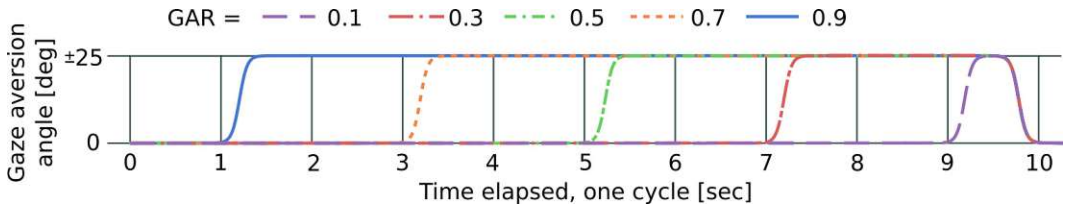

Figure 2.2: Timing of gaze focus transitions for the five conditions. At 0° Pepper looks at the human. The movement time of 0.5s is included in the gaze aversion time. At 9.5 s, all GAR conditions return to 0°. Pepper's different gaze angles are -25°, 0°, and +25°.

eye blinking and gazing, opening and closing grippers, and arm movements. For our experimental design, we iteratively arrived at the following animacy behavior: eye blinking behavior through Pepper's LED eyes, "breathing" behavior through subtle body movements, and short, randomized head movements that are added to the explicit gaze shifts. Moreover, for the experiment, homogeneous interaction length among the participants was desirable. We conducted several HHI pre-trials where a human played the role of the robot. The participants rated a duration of two to three minutes as a comfortable interaction length.

During the greeting and farewell part of the interaction, the robot performed several utterances and gestures (see Appendix A.1). If a participant stopped talking for more than three seconds, the robot made an utterance to ask for more input from the user ("Was that all?") without an additional gesture. The experimenter triggered the beginning of the experiment, i.e., the robot greeting, the robot utterance after a human speaking pause, and the robot farewell utterance manually in the Wizard-of-Oz (WoZ) style [67]. In the pilot study [43], we conducted the main experiment without eye-tracking and only three GAR scenarios 0.1, 0.5, and 0.9 ($n = 10$). These trials were used for wizarding training, and the first author acted as the only wizard for all experiments. The wizard knew the research questions of the experiment, but there were no hypotheses as to which group comparisons were likely to be significant. With respect to the guidelines elaborated in [68], the only wizard recognition and production variables were detecting speech pauses of three seconds and then triggering the next phase of the experiment. If the duration was less than two minutes, the robot asked "Was that all?" to encourage the participants to talk longer. When a three-second speech pause occurred after two minutes or after the robot had already asked to elaborate once, the farewell utterance was triggered.

The microphone integrated into the Pepper robot was not reliable enough for the automatic detection of human speech. In other HRI experiments, this difficulty was solved by providing a hand-held or head-mounted external microphone to the human participant, e.g., [69]. In our experiment, however, this would have made the experiment for the participants too cumbersome, as they already wore a head-mounted eye-tracker. Information about the degree of wizarding was given to the participants only after they filled in the questionnaire. [67] argue that a clear specification of the scenario for the participants is important as well. This was achieved by providing a written instruction of the experiment before consenting, and another verbal instruction provided

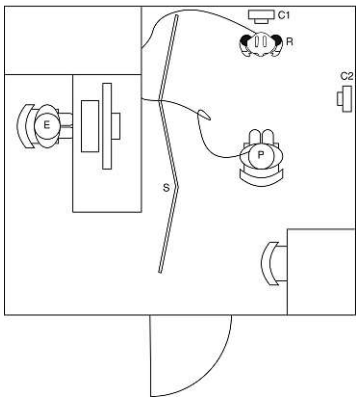

Figure 2.3: Schematic layout of the experiment room ($4m \times 3.5m$). Participant with eye tracker (P), experimenter (E), robot (R), cameras (C1, C2), room separator (S).

by the experimenter. Concerning the eye-tracking procedure, the following steps were performed: 1) explanation of the eye-tracking hardware, 2) fitting of the eye-tracker, 3) adjusting the eye-tracking camera for robust pupil detection, 4) performing the single marker calibration choreography, 5) confirming eye-tracking quality, and 5) recording. Then the robot started the interaction with a greeting, which the experimenter triggered (see Appendix A.1). The experiment design and data processing procedure were peer-reviewed by the TU Wien Pilot Research Ethics Committee.

For the main experiment, all participants ($n = 101$) were recruited from the premises of TU Wien on the fly. Most of them were students. The participants were asked to give written consent after reading a short introduction to the experiment. After the experiment each participant received 10 € as compensation. We performed the experiment in a room of the TU Wien library. To achieve consistent lighting conditions, the room was only illuminated with artificial ceiling lights while the window shades were closed. There was no sound, except the fans of the robot and the computer the experimenter used.

The room was split in two sides by an opaque room separator (Figure 2.3). On one side, the participant sat on a chair in front of the Pepper robot at a distance of 110 cm between the chair and the base of the robot. Two cameras were positioned near the robot and the participant. On the other side of the separator, the experimenter executed the robot interaction script from a desktop. The WoZ controller was able to see the camera feed of the pepper robot and listen to the participant.

COVID-19 hygiene measures were taken in all sessions. Pencils and the eye-tracking device were disinfected before every trial, and the room was ventilated between trials. The experimenter wore an FFP2/N95 mask at all times, while the participants were asked to remove their mask to accommodate the eye-tracker.

To measure behavioral differences between GAR conditions, we record the eye gaze behavior, the interaction duration, and the word count during the interaction as objective measures. To measure attitudes of the participants toward the robot, we compiled sections from validated questionnaire series, namely Godspeed [70] and

BEHAVE-II [71]. We chose scales that are appropriate for the interaction setting and research questions. Each section consists of four 5-point Likert-type scales, each with two positively and two negatively connoted terms, with an additional "I don't know" option. The aggregated scale of *attention* consisted of the ratings *responsive*, *interactive*, *ignorant*, and *unconscious*. The aggregated scale of *comfort* consisted of the ratings *creepy*, *feeling nervous*, *warm*, and *pleasant*. The aggregated scale of *interaction capabilities* consisted of the ratings *artificial*, *incompetent*, *intelligent*, and *sensible* (Appendix A.5). To avoid misunderstandings, we provided the participants with a glossary for the terms of the questionnaire (Appendix A.6).

Additionally, we posed three open-ended interview questions after participants completed the questionnaire to detect problems during the experiment and to get qualitative impressions: "How did it feel to talk to the robot?", "Do you have additional thoughts about the robot?", and "Do you have any other thoughts about the experiment?".

The resulting dataset is securely and confidentially stored at the TU Wien research data repository (doi:10.48436/frswc-4dn44).

2.4 Results

The data analysis is performed for *human gaze behavior* (i.e., gaze fixations on a specific Region of Interest (ROI) and pupil measures), *interaction behavior* (i.e., word count and duration), and *attitudinal measures* (i.e., self-reported measures). The human gaze data is analyzed in different ways: 1) aggregated statistical measures of the whole interaction, 2) aggregated statistical measures for each interaction, 3) human gaze sequences, and 4) temporal correlation of robot and human gaze shifts.

2.4.1 Data Preparation and Descriptive Statistics

A total of 101 participants took part in the experiment. Of these 101 participants, 4 had to be excluded due to technical problems with the experimental procedure, and 1 participant due to misunderstanding the instructions, resulting in a sample size of 96 for the self-reported measures and interaction behavior measures of word count and duration (n=96, Age: mean = 24.26, SD=4.02). A total of 43 participants identified as female and 53 as male. Every participant reported daily computer use. Regarding experience with robots, the participants either never (61), once (12), or a few times (23) interacted with a robot before. No one has interacted with robots on a regular basis until the time of the experiment.

For the analysis, the negatively formulated scales *ignorant*, *unconscious*, *creepy*, *nervous*, *artificial*, and *incompetent* were inverted. Items answered with "I don't know" were treated as missing values and occurred with the following frequencies: *responsive*: 1, *interactive*: 2, *ignorant*: 4, *unconscious*: 5, *creepy*: 0, *feeling nervous*: 0, *warm*: 0, *pleasant*: 0, *artificial*: 0, *incompetent*: 5, *intelligent*: 6, *sensible*: 3.

The aggregated scale of *attention* is composed of the items *responsive*, *interactive*, *ignorant* (inverted), *unconscious* (inverted). The composite scale of *comfort* is composed of the items *creepy* (inverted), *nervous* (inverted), *warm*, *pleasant*. The composite

GAR	0.1	0.3	0.5	0.7	0.9
n	19	19	21	17	20
Age mean (SD)	26.7 (6.4)	23.4 (2.6)	23.6 (2.5)	22.8 (2.2)	24.3 (3.8)
Gender	9 f/ 10 m	8 f/ 11 m	10 f/ 11m	8 f/ 9 m	8 f/ 12 m
n	19	17	19	16	17
Age mean (SD)	26.7 (6.4)	23.4 (2.6)	23.4 (2.4)	23.1 (2.5)	24.6 (3.8)
Gender	9 f/ 10 m	6 f/ 11 m	9 f/ 10m	7 f/ 9 m	6 f/ 11 m

Table 2.1: Top: Descriptive statistics of participants ($n = 96$) for attitudinal and interaction behavior measures. Bottom: Descriptive statistics of participants ($n = 88$) for gaze-related measures.

scale of *capability* is composed of the items *artificial* (inverted), *incompetent* (inverted), *intelligent*, *sensible*. Cronbach's alpha tests for the aggregated scales *attention* ($\alpha = 0.84$), *comfort* ($\alpha = 0.62$), and *interaction capabilities* ($\alpha = 0.73$) were performed.

For the attitudinal and behavioral measure tests, descriptive statistics per test condition are presented in the top of Table 2.1. Discrepancies in the group sizes arose because we also tried to have similar group sizes for valid eye-tracking data, where additional errors occurred that led to the exclusion of some participants.

For the eye-tracking analysis, from the 96 participants who completed the self-report questionnaire, 8 more needed to be excluded due to poor average pupil detection confidence (< 0.6) (n = 88, Age: mean = 24.22, SD = 4.07, Gender: 51 m, 37 f) (bottom of Table 2.1).

Gaze data during the interaction period were recorded with a Pupil Labs Core eye-tracking device. The gaze data was aggregated as follows: From the gaze data per recorded frame, fixations were detected through the Pupil Labs fixation detection algorithm[5]. Thus, each fixation is described by a duration and a gaze coordinate, which can be mapped into world space and, more specifically, onto a specific ROI in the world space. On the most granular level, the following ROIs are defined, as depicted in Figure 2.4: Robot head (H), robot body (B), top left of head (TLH), top of head (TH), top right of head (TRH), left of head (LH), right of head (RH), top left of body (TLB), top right of body (TRB), bottom left of body (BLB), bottom right of body (BRB), bottom left (BotL), bottom (Bot), bottom right (BotR). Gaze at the head is defined as gaze at ROI *head*, gaze at the robot body is congruent with ROI *body*, whereas the gaze at any of the remaining ROIs is coded as ROI *gaze averted*. These individual ROIs constituting the ROI *gaze averted* can further be grouped to distinguish between fixations on *top* and *bottom*.

Gaze fixations that do not fall into any of the defined ROIs are coded with *no ROI*. These fixations have been added to the general *gaze averted* ROI, but have not been used when splitting gaze aversions into *top* and *bottom*. The relevant time span of an interaction started at the end of the greeting utterance of the robot up to the start of

[5]The chosen parameter settings for the fixation classifier are: dispersion duration, maximum dispersion = 3°, minimum duration = 100 ms, maximum duration = 4000 ms, single marker calibration choreography in Pupil Player v3.3.0.

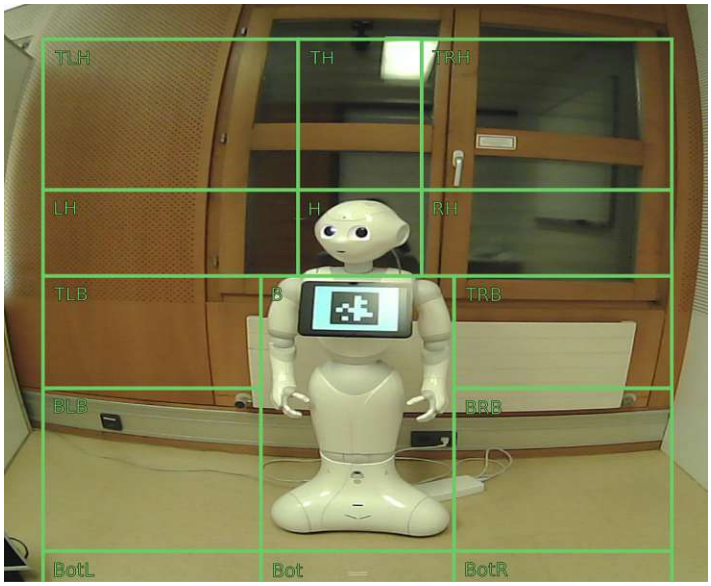

Figure 2.4: Regions of interest (ROI) relative to the Pepper robot as seen from the world view camera of the eye-tracking device from top left to bottom right: Top left head (TLH), top head (TH), top right head (TRH), left head (LH), head (H), right head (RH), top left body (TLB), body (B), top right body (TRB), bottom left body (BLB), bottom right body (BRB), bottom left (BotL), bottom (Bot), bottom right (BotR).

the farewell message or the utterance of the invitation to speak more.

2.4.2 Power Analysis

G*Power [72] was used to compute the power of the study design and the compromise of α and power. For a sufficiently sized sample, the preferred test statistic would be the one-way ANOVA, since normal distribution is expected in the dependent variables. Therefore we computed the a-priori sample sizes using *Cohen's f* as effect size ($f = 0.1$: small, $f = 0.25$: medium, $f = 0.4$: large). With $\alpha = 0.05$, power $(1 - \beta) = 0.8$, and 5 groups, the required sample sizes are $n = 80$ for a large effect, and $n = 200$ for a medium effect. The actual sample size of $n = 100$ is suitable for an effect size of $f = 0.355$.

The used η^2 effect size can be transformed into f using $f = \sqrt{\eta^2/(1-\eta^2)}$ [73], which results in $\eta^2 = 0.0828$, $f = 0.3$ for the *sensible* item in the self-reported data.

A compromise power analysis of $q = \beta/\alpha = 4$ (since $\alpha = 0.5$, and test power $1 - \beta = 0.8$), $N = 100$ and 5 groups resulted in a compromise value of $\alpha = 0.10$ and test power of $(1 - \beta) = 0.60$. With this insight, the hypothesis test results that fall into the range of $0.05 < \alpha < 0.1$ can serve as indicators for future studies, e.g., the capability scale in the self-reported data analysis.

2.4.3 Effect of Gender

We did a preliminary check on the effect of gender, as gender has been reported to have an effect on eye gaze behavior (e.g., [74]). Since the t-test assumptions for a t-test between *attention*, *comfort*, *capability*, *duration*, and *word count* as dependent variable and *gender (f/m)* as independent variable were not met, we performed a Mann-Whitney-U test on these dependent variables, without any significant results. The same applies to the main dependent eye gaze variables, namely the normalized summed-up fixation durations of the ROIs *head*, *body*, and *gaze averted*. These results, together with the gender stratification in the GAR conditions, lead us to exclude gender as a covariate from further analysis.

2.4.4 "Was that all?" - Robot Asked For More Information

Among the 96 participants, the robot asked 44 of them to continue talking about their chosen movie because otherwise, the interaction duration would have been shorter than two minutes; the other 52 participants were not asked to elaborate further. This does not influence the eye gaze measures, since the gaze behavior is only evaluated for the interval between the end of the robot greeting utterance and the next robot utterance after that, i.e., query for more information or farewell. However, the participants filled in the questionnaire after the end of the interaction. Thus, the possibility that the additional question posed by the robot influenced the perception of the participants must be investigated.

We tested for significant group differences for the attitudinal and behavioral measures using a Kruskal-Wallis test (Appendix A.2.1 and A.2.2). There were no significant differences for the self-report data. Thus, we conclude that the additional question by the robot did not influence the attitudinal measures. However, as expected, there were significant differences in the interaction duration and word count. Participants, who were queried for more information, talked for a shorter amount of time.

2.4.5 Human Gaze Data Evaluation

The 88 participants with valid eye-tracking produced a total of 23529 fixations. The average fixation count per participant was 247 (SD = 158). The average fixation duration per participant was mean = 404 ms (SD = 215 ms). In total, 913 fixations were categorized as *no ROI*. The descriptive statistics per GAR condition are shown in Appendix A.3.3.

2.4.5.1 Fixation Duration

We investigate whether the fixation durations differ between the groups per ROI, when aggregated over the whole interaction. We normalized the fixation duration sums per participant and averaged them per condition for the three ROIs *head*, *body*, and *gaze averted* between the GAR conditions. The ANOVA test assumptions are met for the

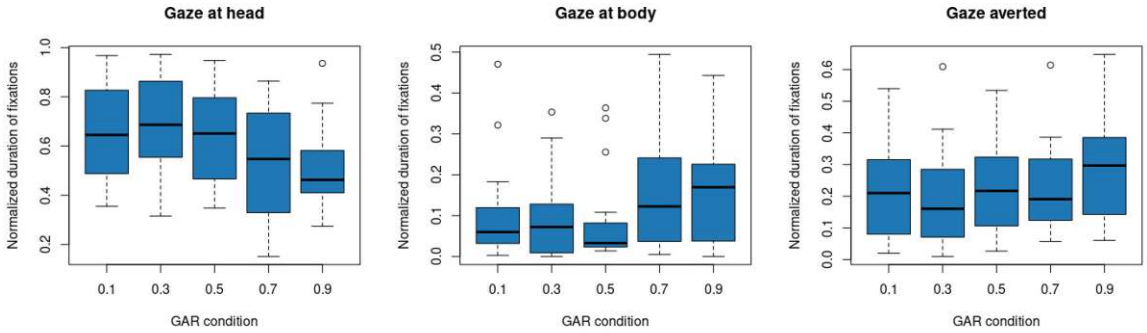

Figure 2.5: Normalized gaze duration on ROIs *head*, *body*, *gaze averted* for each GAR condition.

gaze duration at the head but not for the body and the averted ROI.[6]

We performed an ANOVA on the *head* ROI (Appendix A.3.4) between GAR conditions ($F(4) = 2.704$, $p = 0.036, \eta^2 = 0.117$). We performed a Kruskal-Wallis test on the ROI *body* ($\chi^2(4) = 6.4074, p = 0.1707, \eta^2 = 0.0294$) and the ROI *gaze averted* ($\chi^2(4) = 3.15, p = 0.533, \eta^2 = -0.0104$). These results indicate that the GAR has an effect on ROI head but not body and gaze averted. However, a Tukey HSD post-hoc test to find pairwise differences in the fixation duration of the *head* ROI revealed no significant group differences. The lowest adjusted p-value occurred between the two groups GAR 0.3 and 0.9 (p = 0.065).

Figure 2.5 motivated a correlation and linear regression analysis between GAR conditions and fixation duration on ROI *head*. GAR is an interval scale regarding its implementation on the robot. We performed a Pearson correlation ($r = -0.2981547$, $t(85) = -2.8798$, $p = 0.005$). The linear regression resulted in an intercept of 0.71 (Std.Err = 0.04, $t = 16.98$, $p < 0.000$) and slope (i.e., the effect of the condition) of -0.21 (Std.Err= $0.07, t = -2.88, p = 0.005$). The multiple R-squared value of the linear regression is 0.089. Visual inspection of the residuals (Figure 2.6) does not suggest higher order functional relations between GAR and fixation duration at ROI *head* and also, a quadratic regression analysis was non-significant.

The regression and correlation indicate that the lower the robot GAR (i.e., the more the robot gazes at the human) the more the human gazes at the face of the robot. Conversely, the more the robot gazes away from the human, the more the human gazes at the body of the robot.

2.4.5.2 Fixation Duration - Temporal Split

Until now, we inspected gaze metrics that are temporally aggregated across the whole interaction duration. In this section, we analyze the influence of time on gaze behavior in the different GAR conditions. Therefore, we split each individual interaction into thirds, similar to [57] and [75].

[6]The results of the analysis of the fixation counts yield largely the same results, but using the summed fixation duration is more widely used and more principled.

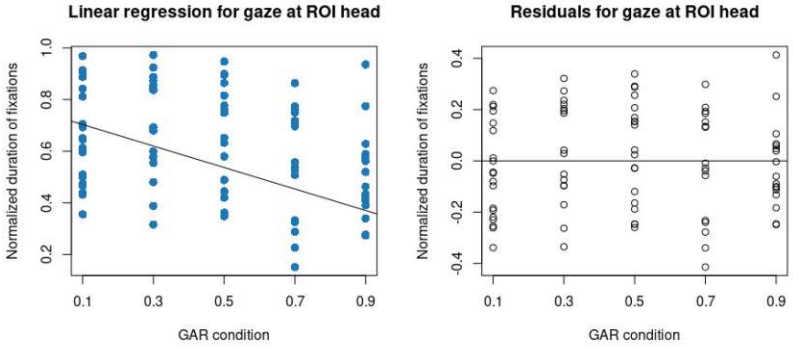

Figure 2.6: Left: Linear regression model between GAR and fixation duration at ROI *head*. Right: The residuals of the linear regression.

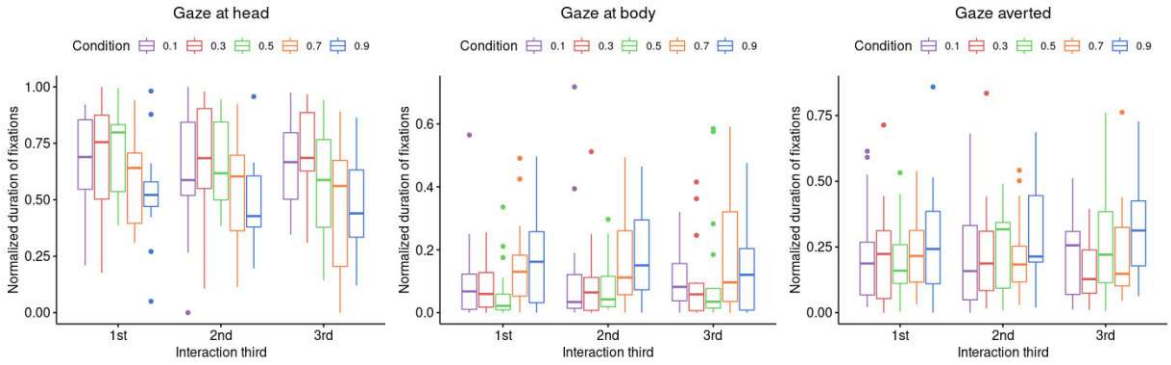

Figure 2.7: Normalized fixation durations for variable ROI *head* (left), *body* (middle), and *gaze averted* (right) for the between-factor *GAR condition* and the within-factor *interaction third*.

For each third, the gaze metrics are calculated. The ANOVA assumptions are met for the ROI *head*, but not for the other two ROIs *body* and *gaze averted*. Thus, we performed a MANOVA on the ROI *head* with the within-factor *time* and the between-factor *condition* (Table 2.2, Figure 2.7), which revealed significant main effects, as well as a significant interaction effect, all with a small effect size.

To determine where the interaction effect lies, we performed two groups of one-way ANOVAs: First, we examined the effect of the condition in the separate interaction thirds (middle of Table 2.2). The condition has a significant and large effect in the last interaction third.

Second, we examined the effect of time in the different conditions via another group of ANOVAs (bottom of Table 2.2). There was a significant difference in the time spent gazing at the robot head (medium effect) between the interaction thirds in the GAR condition 0.5. This finding emphasizes the visual result of Figure 2.7, where the gaze at the head in the 0.5 GAR condition occurs for a longer duration in the first third than in the other two thirds. Summarizing the interaction effect, an inspection of the median values in Figure 2.7 suggests that the gaze durations form a "∧" shape in the first third, whereas this is transformed into a downward slope "

	F	df1	df2	p	p.adj	η^2
Condition	2.54	4	80	0.046*		0.093
Time	4.64	2	160	0.011*		0.011
Interaction	2.13	8	160	0.035*		0.020
Time	F	df1	df2	p	p.adj	η^2
1st third	1.98	4	82	0.106	0.110	0.088
2nd third	2.42	4	82	0.055	0.112	0.105
3rd third	3.38	4	80	0.013	0.039*	0.145
Condition	F	df1	df2	p	p.adj	η^2
0.1	0.512	2	36	0.604	1.000	0.007
0.3	0.985	2	28	0.386	1.000	0.016
0.5	12.781	2	34	<0.000	<0.000*	0.070
0.7	4.424	2	30	0.021	0.084	0.043
0.9	0.348	2	23	1.000	1.000	0.017

Table 2.2: Top: MANOVA for the within-factor *Time* and the between-factor *condition* ($\alpha = 0.05(*)$). Middle: ANOVA with the factor *condition* on the normalized fixation duration of the ROI *head* for each interaction third. Bottom: ANOVA with the factor *Time* on the normalized fixation duration of the ROI *head* for each condition.

" in the last third over time.

Finally, for the ROI *head*, Tukey HSD post-hoc tests for the interaction effect revealed no significant comparisons, however this can be explained by the unusually high number of group comparisons when both *interaction third* and *condition* are examined. For the other two ROIs *body* and *gaze averted*, the ANOVA assumptions were not met. Therefore the multivariate non-parametric Scheirer-Ray-Hare test was chosen and performed on ROI *body* (GAR condition: H(4) = 16.4, p = 0.002, time: H(2) = 0.55, p = 0.76, interaction H(8) = 3.60, p = 0.89) (Figure 2.7). Only the GAR condition had a significant effect, and Dunn post-hoc tests revealed the two different GAR groups (GAR 0.3, 0.5 and GAR 0.7, 0.9), with the second group exhibiting a higher amount of gazing at the body. GAR 0.1 almost met the significance level for being significantly different from GAR 0.7 and 0.9. The significant group comparisons are 0.1-0.7 (adjusted p (p adj.) = 0.06), 0.1-0.9 (p adj. = 0.06), 0.3-0.7 (p adj. = 0.03), 0.3-0.9 (p adj. = 0.02), 0.5-0.7 (p adj. = 0.02), and 0.5-0.9 (p adj. = 0.01). This reinforces the linear regression result of the ROI *head*: People looked at the body of the robot - and not at the robot face - more often when the robot displayed a high GAR.

The procedure was repeated for ROI *gaze averted* (GAR condition: H(4) = 8.79, p = 0.06, time: H(2) = 0.50, p = 0.77, interaction H(8) = 2.98, p = 0.93) (see Figure 2.7) without any significant effects.

2.4.5.3 Analysis of Gaze Sequences

Similar to [55], [62], [76] we are interested in whether the different GAR conditions lead to different dynamic gaze patterns in the participants. The gaze directions of the human participants are categorized into ROIs (Figure 2.4). Thus, every participant produced a sequence of fixations on the ROIs, similar to [77] and [78]. For this section, we group the ROIs into the following gaze directions (Figure 2.4): Head ($H = \{H\}$), Body ($B = \{B\}$), Up ($U = \{TH\}$), Up-Left ($UL = \{TLH, LH\}$), Down-Left($DL = \{TLB, BLB\}$), Up-Right ($UR = \{TRH, RH\}$), Down-Right ($DR = \{TRB, BRB\}$), Down ($D = \{BotL, Bot, BotR\}$). Therefore each participant $i, (i \in \{1, \ldots, N\})$ produces a sequence $S_i = \{S_{i1}, S_{i2}..., S_{iT_i}\}$, with $S_{it} \in D = \{H, B, U, UL, DL, UR, DR, D\}$ and $t \in \{1, \ldots, T_i\}$, with T_i as the length of the fixation sequence of participant i. For each participant, a sequence of *fixation transitions* is constructed. They can be used to create a stochastic model of the gaze behavior of a single participant in the form of a Discrete-Time Markov Chain (DTMC) [79]. For these models, the Markov property assumption dictates that the gaze direction probability depends only on the previous gaze direction. This assumption is expressed as

$$Pr(S_{t+1} = s_{t+1} \mid S_1 = s_1, S_2 = s_2, \ldots, S_t = s_t) = Pr(S_{t+1} = s_{t+1} \mid S_t = s_t), \quad (2.1)$$

where the single fixation directions are random variables with the domain D, thus $s_t \in D$. Additionally assuming time-invariance across an interaction, the stationary distribution is represented as a $|D| \times |D|$ transition matrix M, where an element m_{ij}, $i, j \in \{1, \ldots, |D|\}$, describes the probability of transitioning from direction i to direction j, with $\sum_{j \in \{1,\ldots,|D|\}} m_{ij} = 1$ for all $i \in \{1, \ldots, |D|\}$. Transitions include loops, which result from starting and ending in the same ROI. This can occur, when a participant has two consecutive fixations in the same ROI.

Subsequently, the transition matrix for each participant is created. To aggregate them, the single participant transition matrices are summed up and normalized row-wise. This way, the gaze behavior of each participant influences the outcome with the same weighting, independent of the interaction duration.

For each GAR condition, there is a different DTMC model. Each DTMC is a 14×14 transition matrix. Visual inspection of the differences between these matrices can reveal how the gaze transitions differ. However, to determine whether the models differ from each other in a statistically significant way, each of the five 14×14 matrices is flattened into a vector of length 196. Then, these five vectors can be stacked into a 5×196 matrix. In this representation, each column describes one fixation transition, e.g., *head* $\rightarrow$ *body*. Considering the non-normalized fixation counts, each row describes the categorical distribution of fixation shifts. This matrix is very sparse, containing many zero values for unlikely transitions. To interpret the results and to perform a χ^2 test of independence, we aggregated the fixation shifts into the transitions between the previously defined ROI regions *head*, *body*, and *gaze averted*. This transformation yields a 5×9 matrix, with all cell values > 5 (Table 2.3).

For this matrix we performed a χ^2 test of independence ($\chi^2(32) = 973.57, p < 2.2e{-}16$, Cramer's V = 0.108 (weak effect)). To determine which cells lead to this significant

GAR	b-b	b-h	b-a	h-b	h-h	h-a	a-b	a-h	a-a
0.1	205	153	67	160	1870	313	56	324	488
0.3	225	161	79	161	1517	380	77	374	585
0.5	249	150	67	144	2298	537	77	530	1116
0.7	481	248	131	268	1322	394	103	409	749
0.9	510	267	133	267	1263	401	134	378	941
0.1	-87***	−18	−17	−15	422***	−42	−22	−29	-191***
0.3	-61**	−7	−3	−10	99**	33	0	29	-80**
0.5	-167***	-94***	-52***	-105***	239***	33	-34**	28	150***
0.7	151***	54***	37***	70***	-313***	−6	15	11	−18
0.9	165***	65***	34**	60***	-447***	−18	42***	−39	139***

Table 2.3: Top: Observed transition counts between the ROIs *body* (b), *head* (h), and *gaze averted* (a). Bottom: Differences of transition counts (observed - expected) for the transitions between the ROIs *body* (b), *head* (h), and *gaze averted* (a) per GAR condition. The adjusted α level is $0.00\dot{1}$ (*), $0.000\dot{2}$ (**), and $0.0000\dot{2}$ (***). Colored cells highlight comparisons made in section 2.4.

result, the z-scores for the cells were calculated (where a $z < -1.96$ or $z > 1.96$ would represent a significant deviation without Type I error correction). Then, the χ^2 values were calculated using the z-scores. For each one of the χ^2 values, the p-value was compared against a χ^2 distribution (df = 1, one-sided) [80], [81]. The adjusted significance level $\alpha = 0.5/(5 \cdot 9) = 0.00\dot{1}$ was used to check which cells differ significantly from the expected distribution (Table 2.3). For example, all entries in the body-to-body transitions column are significant. Negative cell entries indicate that less gaze transitions than expected have occurred, while positive cell entries indicate that more transitions than expected occurred. Figure 2.8 visualizes all significant deviations from expected gaze transition frequencies. In that figure, blue solid arrows indicate a significant positive result (i.e., more transitions than expected), while orange dashed lines indicate significant negative results (i.e., fewer transitions than expected). This graphic only indicates significant deviations from expectation and not the absolute number of transitions. For example, comparing the two columns *h-b* and *h-a* (marked in red color) in Table 2.3 reveals that *h-a* transitions occurred more frequently than *h-b* transitions among all groups. This means, in general after gazing at the head, participants averted their gaze, rather than look at the robot's body.

However, we are interested in differences between the GAR conditions, and thus we proceed to cluster the significant cell entries in Table 2.3 to interpret them on a higher level. The three self-loop columns in Table 2.3 for *h-h*, *b-b*, and *a-a* (cells colored in blue) confirm the findings of the gaze behavior with respect to the total time spent gazing at a particular ROI. All entries in column *h-h* display significant differences from the expected occurrences, namely a higher count for the low GAR conditions 0.1, 0.3, and 0.5, and a lower than expected count for the conditions 0.7 and 0.9. Comparable results are visible in the columns *b-b* and *a-a*, where the relationship is reversed for the GAR conditions. This means that when the robot gazed at the participants for a longer

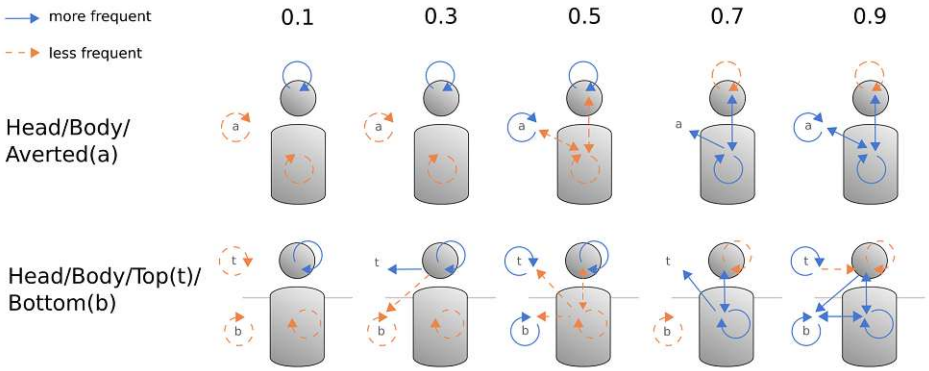

Figure 2.8: Visualization of significant deviations from the expected gaze transition frequencies in the 5 GAR conditions. Top: ROIs *head*, *body*, *gaze averted*. Bottom: ROIs *head*, *body*, *top*, *bottom*.

time, the participants reciprocated the robot's mutual gaze.

Regarding true fixation shifts, i.e., transitions between different ROIs, a large pattern of significant results can be observed in the submatrix GAR 0.5, 0.7, 0.9 and transition *b-h*, *h-b*, *b-a*, and *a-b* (cells colored in orange). These transitions occur less frequently for the 0.5 GAR condition and more frequently for the GAR conditions 0.7 and 0.9. This means, in the 0.5 GAR condition, participants shifted their gaze less frequently between the body and the head, and less frequently between the body and the *gaze averted* ROI (while in this GAR condition, gazing at the head was more frequent). For the conditions 0.7 and 0.9 the relationship is reversed: There is more fixation shifting between head and body, with less frequent gazing at the head. Similarly, there are more transitions between the body to the *gaze averted* ROI. This means, the gaze pattern in the 0.7 and 0.9 conditions centers more around the body of the robot, while in the 0.5 condition, gaze at the body occurs less frequently. In short, in the 0.5 GAR condition there occurred significantly fewer fixation shifts between different ROIs, while in condition 0.7 and 0.9, there occurred significantly more fixation shifts between different ROIs.

The gaze transition *a-b* happened less frequently in the 0.5 condition and more frequently in the 0.9 condition. Interestingly, there are no differences for the gaze shifts *h-a* and *a-h*. This means that there is no difference between the GAR conditions for looking at the robot's head after averting the gaze, however, this is the gaze shift the robot performed.

In summary, these results confirm that robotic GAR influences human gaze behavior, but not in a direct way. Otherwise, the frequency deviations in the gaze shifts between the ROIs *head* and *gaze averted* would be significant since the robot gaze aversion pattern consists solely of shifts between the ROIs *head* and *gaze averted*. Participants reacted with more gaze aversion to a high gaze aversion pattern of the robot. However, the participants seemed not to perform mere gaze following. Deviations from the expected frequencies occurred in the gaze pattern that centered around the robot body to then shift the gaze to the robot head or averting the gaze.

Gaze aversions in different directions have been associated with different conversational goals. [23] report that people in conversations more often avert their gaze upwards

GAR	b-t	b-bot	h-t	h-bot	t-b	t-h	t-t	t-bot	bot-b	bot-h	bot-t	bot-bot
0.1	22	45	254	59	19	263	369	24	37	61	30	65
0.3	27	52	353	27	35	336	454	27	42	38	24	80
0.5	17	50	467	70	23	447	824	41	54	83	27	224
0.7	56	75	343	51	38	346	595	39	65	63	36	79
0.9	35	98	318	83	41	294	688	28	93	84	25	200
0.1	−5	−11	−50	8	−8	−32	-144***	−4	−14	3	5	-48***
0.3	0	−3	56*	−23*	8	47	−48	0	−8	−18	0	−31
0.5	−22**	−30**	35	−2	−16	27	95***	1	−18	1	−8	63***
0.7	25***	12	0	−6	7	13	16	8	7	−2	8	-49***
0.9	3	32***	−41	23*	9	-55*	82**	−5	33***	16	−4	66***

Table 2.4: Top: Observed transition counts between the ROIs *body* (b), *head* (h), *top* (t), and *bottom* (bot). Bottom: Differences transition count (observed - expected) for the transitions between the ROIs textitbody (b), *head* (h), *top* (t), and *bottom* (bot). The adjusted α level is 0.00083(*), 0.00017(**), and 0.000017(***). The transitions *h-h, h-b, b-h, b-b* are omitted since they are already depicted in Table 2.3. Colored cells highlight comparisons made in section 2.4.

when they experience a high cognitive load and they gaze downwards when they need to regulate the level of intimacy. Therefore, we split the single ROIs of the aggregated ROI *gaze averted* into *top* (*t*) and *bottom* (*bot*). The ROI *top* consists of all ROIs above *body*, except the ROI *head*. The ROI *bottom* consists of all ROIs below the *head* except *body* (Figure 2.8). The χ^2 tests of independence was significant ($\chi^2(60) = 1122.2, p < 2.2e-16$, Cramer's V = 0.116 (weak effect)). The results for single significant cells are shown in Table 2.4. We adjusted the α level for *top-bottom* to $\alpha = 0.5/(5 * 12) = 0.0083$.

Concerning the general gaze pattern for all groups, *head* to *top*, *top* to *head*, and *top* to *top* gaze shifts were far more numerous than the gaze shifts regarding the ROI *bottom* and body (cells marked in blue). This gaze aversion pattern rather indicates a cognitive effort than intimacy regulation. Checking the top to bottom and body to bottom columns reveals that this cannot be a measurement artifact. Assuming that the true gaze shift occurred between head and bottom, but the eye tracker falsely registered an intermittent fixation on top or body, we would expect more shifts from top to bottom or body to bottom column. However, the frequencies in both columns are relatively low, too.

As mentioned above, we are, however, interested in group differences between the GAR conditions and trying to find an explanatory pattern for the significant cells. The following significant results were observed: Participants in the 0.5 and 0.9 conditions had more fixation shifts from *bottom* to *bottom*. In contrast, in conditions 0.1 and 0.7, this occurred less frequently (cells marked in red). The *top* to *top* transitions occurred less frequently in the 0.1 condition and more frequently in the 0.5 and 0.9 condition (cells marked in orange). There were no other significant results for the 0.1 condition. Participants in condition 0.3 made more gaze shifts from *head* to *top* than from *head* to

bottom. In condition 0.5, participants gazed less frequently from *body* to *top* and from *body* to *bottom*. In condition 0.7, more gaze shifts occurred from *body* to *top*. There were more significant results in condition 0.9: There were more occurrences of shifts from *body* to *bottom*, *head* to *bottom*, and *bottom* to *body*, and fewer occurrences of *top* to *head* (cells colored in blue).

Top to *bottom* or *bottom* to *top* fixation shift frequencies did not differ between groups. Looking for a pattern reveals significant results in the *bot* - *bot* column, which are largely congruent with the *a-a* gaze shift in the previous ROI split. For condition GAR 0.9, the gaze shifts paint a picture of higher frequency gazing downwards, which can be indicative of intimacy regulation [23] in contrast to the other groups. In summary, splitting the *gaze averted* ROI into *top* and *bottom* reproduces the previous result.

2.4.5.4 Temporal Correlation of Robot and Human Gaze Shifts

Next, we wanted to examine whether the gaze behavior of the robot causes temporally aligned fixation shifts of the participants. Therefore, for each participant, we calculated the point in time of each real fixation shift (i.e., no loop like head-to-head) with respect to the ten second gaze cycle time of the robot. Thus, each fixation has a starting time between 0 s and 10 s with respect to the gaze cycle duration. These starting times per participant are gathered per GAR condition. This results in five distributions, which we tested for significant differences. Since the variables were not normally distributed, we conducted one Kruskal-Wallis test counting all real fixation shifts ($\chi^2(4) = 5.36, p = 0.25$) and one Kruskal-Wallis test that counted only fixation shifts from and to the head ($\chi^2(4) = 4.31, p = 0.36$). Thus, there is no evidence for differences in the temporal occurrence of gaze shifts within the 10 s robot gaze cycle. The distribution for all GAR conditions is depicted in Figure 2.9. The depicted histograms suggest a uniform distribution within the 10 s cycle for all five conditions. This result indicates that the robot gaze shifts do not trigger human gaze shifts. In the previous sections, differences in the gaze behavior between the GAR conditions were found, but the differences cannot be explained by mere gaze following.

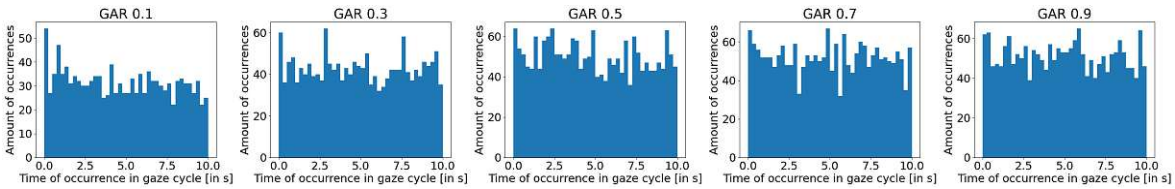

Figure 2.9: The distributions of the occurrence of a fixation shift of the participants with respect to the 10 s gaze cycle of the robot. From left to right: GAR 0.1, 0.3, 0.5, 0.7, 0.9.

2.4.6 Interaction Behavior Evaluation

The two variables *duration* and *word count* (Table 2.5) were naturally highly correlated ($\rho = 0.922$, $t = 23.139, p < 0.000$). The ANOVA assumptions for the effect of GAR on these two variables were not met, therefore we applied a Kruskal-Wallis test (Figure 2.10).

The null hypothesis of RQ 5a, that there is no effect of the robot's GAR on the amount of words spoken by the participants, can be rejected ($\chi^2(4) = 11.3, p = 0.02, \eta^2 = 0.08$). The null hypothesis of RQ 5b, that there is no effect of the robot's GAR on the interaction duration, can be rejected ($\chi^2(4) = 11.6, p = 0.02, \eta^2 = 0.08$). For both significant tests, there is a moderate effect size.

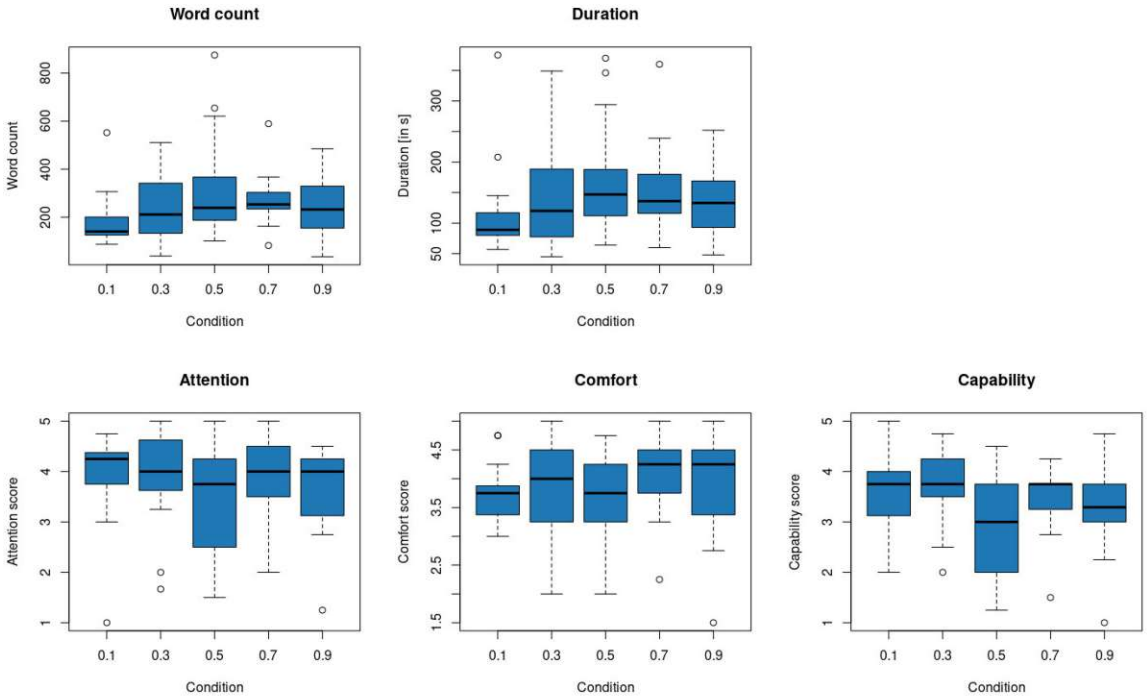

Figure 2.10: Top: Word count (left) and duration (right) per GAR condition. Bottom: Attention displayed by the robot (left), comfort during the interaction (middle), and perceived robot capability (right) per GAR condition.

We conducted pairwise comparisons for *word count* and *duration* using the Dunn test with Bonferroni-Holm correction. There are significant differences in the word count and duration between group 0.1 and 0.5 (p adj = 0.03, p adj = 0.02), as well as 0.1 and 0.7 (p adj. = 0.03 p adj. = 0.04).

This means participants in the GAR 0.1 condition (the robot mostly stares at the participant) spoke fewer words and for a shorter amount of time than in the 0.5 (robot averted gaze half of the time) and 0.7 GAR conditions. Thus, there is a moderately sized effect of the GAR on the interaction level. In conversational settings, listeners tend to gaze at the speaker more often than the speaker gazes at the listener [48]. However, we observe that in the 0.1 GAR condition, participants talked significantly less. A possible explanation for this is that a too low GAR is perhaps not (only) interpreted as benevolently paying attention to the speaker. Together with the gaze behavior result, we observe now that participants in lower GAR conditions (i.e., more robot staring) talked less while gazing more at the robot.

	Word count			Duration			Attention		Comfort		Capability	
Condition	mean	median	SD	mean	median	SD	mean	SD	mean	SD	mean	SD
0.1	180	140	106	113	89	72	3.92	0.85	3.70	0.50	3.59	0.72
0.3	232	211	139	138	120	80	3.94	0.90	3.76	0.81	3.67	0.69
0.5	309	239	197	168	147	82	3.47	1.15	3.74	0.65	2.94	0.93
0.7	272	253	107	157	136	66	3.84	0.82	4.04	0.66	3.44	0.65
0.9	243	232	120	137	133	55	3.70	0.81	3.89	0.87	3.34	0.81

Table 2.5: Descriptive statistics of the *word count* and *interaction duration*, as well as the Likert scale *comfort*.

2.4.7 Attitudinal Data Evaluation

Up until now, only behavioral data were evaluated, which revealed differences in gaze behavior and interaction duration between the GAR groups. In addition to this, self-reported attitudinal data were recorded to learn more about how participants experienced the conversational situation. The preconditions for ANOVA were not met. We performed Kruskal-Wallis tests to evaluate the research questions about the impact of the GAR conditions on the aggregated scales *attention* displayed by the robot, *comfort* elicited by the robot, and *capabilities* of the robot (Appendix A.3.5). The mean and standard deviations of the aggregated scales per condition are shown in Table 2.5 and Figure 2.10.

There are no significant differences between the GAR conditions with respect to the scales *attention*, *comfort*, and *capability* (Table S5). If there is a meaningful effect, it cannot be detected due to the sample size or the actual effect size being too small. Otherwise, the GAR could just not have a significant effect on attitudinal measures. So, although there is a moderately sized effect of the GAR on the behavior level, both for gaze behavior and interaction duration, the participants did not form significantly diverging opinions about the robot in the different conditions.

Additional exploratory evaluations revealed correlations between measures. The Pearson correlation coefficient r was calculated for the three composite self-reported scales *attention*, *comfort*, and *capability* on the one hand and the behavioral data *word count* and *duration* on the other (Table 2.6). In these variables, when aggregated for the whole dataset, no extreme outliers were detected. The data suggests a moderate negative correlation between the participants' perception of attention shown by the robot and both interaction duration and word count. The longer the interaction took, the less the participants felt the robot was paying attention to them. A possible reason is the missing backchannel communication (e.g., utterances like "mhm", "aha") or gestures (e.g., nodding) that are used by humans during prolonged periods of listening.

Correlations for the attitudinal measures and the physiological measures *average pupil size* and *pupil size variance* revealed a significant moderate positive correlation between the pupil size and the attention score (Table 2.6). There is also a significant moderate positive correlation between pupil size variance and level of comfort. These findings affirm previous psychophysiological studies: [82], [83] found that cognitive processes are associated with constantly higher pupil dilations. One can speculate that participants

	Attention			Comfort			Capability		
	t	p	r	t	p	r	t	p	r
Duration	−2.81	0.005**	−0.27	−0.47	0.63	−0.04	−2.00	0.04*	−0.20
Word count	2.72	0.007**	−0.27	−0.18	0.85	−0.01	−2.20	0.02*	−0.22
Avg. pupil size (mm)	2.22	0.02*	0.23	−0.27	0.78	−0.02	−0.47	0.63	−0.05
Pupil size SD (mm)	1.56	0.12	0.16	1.97	0.05*	0.21	−0.33	0.73	−0.03

Table 2.6: Pearson correlations between self-reported (*attention*, *comfort*, and *capability*) and behavioral measures (*duration* and *word count*) and pupil size (*avg. pupil size* and *pupil size SD*) for the whole sample.)

who themselves put more effort into retelling a movie then also exhibited more cognitive effort, which led to a larger pupil dilation and a higher self-worth protecting assignment of the robot's attention to their story. Similarly, participants who might have felt a certain comfort in the interaction thus exhibited a higher pupil dilation variance. This is in accordance with [84], where pupil dilation variance is positively associated with affective processing. If this is the case in this experiment, to arrive at the positive correlation between pupil variance and comfort attribution, the elicited affect spectrum ranges from neutral to positive.

2.5 Discussion

In this section, we will discuss and summarize 1) the findings regarding the three research questions, 2) the limitations of the study, and 3) design recommendations resulting from our work.

2.5.1 Answering the Research Questions

In the previous chapter, the statistical test results of the research questions were presented. For the three subquestions (RQ 4 a-c) on gaze behavior, several different analytical methods were used because of the wealth of data resulting from the eye-tracking procedure. For the two subquestions (RQ 5 a and b) on interaction behavior and the three subquestions (RQ 6 a-c) there are fewer metrics, and they will be used to corroborate the gaze tracking results.

2.5.1.1 RQ 4 - Gaze behavior

For all three subquestions of RQ 4, the null hypothesis can be rejected: GAR does have an effect on the fixation durations (RQ 4a). The effect of GAR on the fixation durations changes during the interaction (RQ 4b). The GAR does influence the fixation sequence behavior of the participants (RQ 4c).

The analysis of the fixation shifts was structured to advance from overall differences of fixation durations on large ROIs *head*, *body*, and *gaze averted* to fine-grained dynamic gaze shift patterns.

The main point of discussion is whether the robotic gaze behavior influences the gaze behavior of the participants and the attitudes of the participants towards the robot. We use gaze as a proxy measure of attitudes towards the robot by incorporating correlations between gaze behavior and affect presented in the related work. These results are compared with the self-reported attitudes and the interaction duration, which is another proxy measure of engagement.

Regarding the linear regression results across all groups for the overall *head* GAR behavior between the groups, there are two competing interpretations. 1) Participants mirror the GAR of the robot, i.e., they create rapport [85] or 2) Participants find the interaction with higher GAR more uncomfortable and therefore avert their gaze more [56]. We discussed another proxy measure from the related work, namely the interaction length [54]: The longer a participant keeps up the conversation, the more engaged they are with the robot and this is a sign for comfort during the interaction. Because participants in the 0.1 GAR condition (high robot mutual gaze) talked for a shorter amount of time, this is an indication that the rapport explanation is more likely.

However, there is a noteworthy deviation from the linear regression: In the linear regression of the normalized fixation durations for the whole interaction, the 0.1 GAR setting does not adhere to the otherwise linear relationship of between the participant and the robot GAR. For a stronger relation, the median of the 0.1 setting was supposed to be higher than the median of the 0.3 setting, but this was not the case. This might indicate that the overall trend of participants mirroring the robot GAR is not valid at the 0.1 end of the GAR spectrum, where the robot stares at the human. Otherwise, when the robot has a higher GAR and thus averts the gaze more often, participants mirror this behavior instead of increasing the gaze towards the robot. If we suppose that the robot was perceived as a social agent and creating rapport is a typical human adaptation in a conversation setting, another effect must have been stronger and at this end of the robot GAR scale and prevented GAR mirroring. This might indicate that, indeed, the 0.1 setting of the robot was perceived as uncomfortable, and participants felt being stared at, for which they sought to regulate the intimacy level by avoiding their own gaze more often. This is an indication that high mutual gaze is not always a justified proxy measure of comfort in the interaction, as previously suggested in [52], [53], [57], [58].

Next, we studied how the fixation distribution changes over the duration of the interaction. Similar to [51] and [58], we also observed a time effect. In their study, the gaze towards the robot head increased over time, whereas [58] observed a decrease in time spent gazing at the robot head. The second result is replicated in our work. [51] argue that later interactions represent a more stable interaction pattern. If this statement was also applicable within single interactions, we can interpret the decreasing gaze at the robot as an effect of habituation or a wearing off of the novelty effect. Regarding the interaction effects, the GAR condition has the strongest effect in the last interaction third, where we can observe a linear regression. However, as for the linear regression of the whole interaction, the median of the 0.1 GAR is not as high as expected. Time as a single factor also showed an effect across all conditions, namely the decrease of time spent gazing at the head of the robot. The strongest interaction effect of time and GAR occurs in the 0.5 GAR condition. The gaze at the robot's head

is higher in the first third than in the other two thirds.

Next, we were interested in the fixation shifts of the participants. The statistical tests revealed which fixation shifts occurred more or less frequently than expected. The fixation shifts that started and ended in the same ROI largely reproduced the findings of the normalized fixation duration analysis: The true fixation shifts, i.e., when the start and end ROI are not equal, show a pattern of higher gaze aversion for the 0.7 and 0.9 GAR setting, but noticeably a pattern of less gaze aversion in the 0.5 condition. This could be indicative of a higher comfort for the participants in the 0.5 condition. However, this identified pattern also does not follow the results of the linear regression mentioned above.

We tried to distinguish between top and bottom gaze aversion. There, the gaze aversion pattern for the 0.7 and 0.9 GAR condition was indicative of an intimacy regulation gaze aversion pattern that typically occurs in uncomfortable situations [23].

Lastly, in this section, we explored the temporal relationship between robot and human gaze shifts, i.e., if the human participants engage in gaze behavior after a certain amount of time after a robot gaze shift. This test produced non-significant results and might be another indicator towards the conclusion that humans react to the robotic gaze behavior, but not by merely gaze following the robot.

2.5.1.2 RQ 5 - Interaction behavior

We can reject the null hypothesis for RQ 5 and thus state that the GAR does indeed have an effect on the amount of spoken words (a) and the interaction duration (b). In the related work, interaction duration was used as a proxy measure for engagement [54]. Using the same argument, we can conclude that in the 0.1 GAR setting, engagement was the lowest. This, together with the conclusion of the 0.1 linear regression outlier in the above subsection, paints a picture of an uncomfortable situation in the 0.1 setting.

This, however is in contrast to the exploratory identified significant negative correlation between *attention* and interaction duration. This might be an indication that interaction duration should not be used as a proxy measure of interest alone.

2.5.1.3 RQ 6 - Attitudes

The results for self-reported attitudinal scales perceived attention shown by the robot (RQ 6a), level of comfort of the participants (RQ 6b), and the perceived robot capability of the robot (RQ 6c) did not significantly differ between GAR conditions. We argue against the existence of a medium or large effect of GAR on the conscious perception of the robot interaction in this interaction setting. In the following, we want to contextualize our line of thought.

The median value for attention was highest in the 0.1 GAR setting (though not significantly different from other settings). Together with the shortest interaction duration, this could be an indication that that participants felt “watched” in a negative sense. The lowest median *attention* and *cabability* value occurred in the 0.5 GAR group, where there was significantly less gaze-shifting behavior on the body of the robot, though these values were not significantly different.

In the qualitative interview, participants repeatedly mentioned that they experienced the robot GAR as a "turning the ear towards the speaker to better listen". Most participants mentioned both positive and negative aspects of the interaction, e.g., "It was weird in the beginning, then it was nice.", "jerky moves, but cool and nice", "less weird than expected", "funny and strange". The robot implementation resulted in a range of descriptions, even for the same GAR setting: "The robot seems alive, especially the eye blinking and head movements." on the one hand, and "The robot did not react at all, there was no interaction." on the other hand. Both statements occurred in the 0.1 GAR setting. Many participants mentioned a missing backchannel communication of the robot while listening.

The occurrences of "I don't know." answers in the attitudinal measures were counted as omissions and are an indicator as to what participants can judge with certainty. Participants always knew whether the robot was *creepy*, they were *feeling nervous*, the robot appeared *warm*, *pleasant*, or *artificial*, all with zero omissions. However, they were less certain about whether the robot was *ignorant*,*unconscious*, *incompetent*, or *intelligent*, all with four to six omissions. This might be an indication that the type of data collection is inadequate for certain topics. The participants were sure when introspective questions were posed and more unsure when statements about the inner working of the robot were asked about. More specific questions tailored to the interaction might be adequate for such external evaluations.

2.5.2 Limitations

The following limitations and considerations should inform future work. The remarks concern the interaction setting and the statistical evaluation.

Participants were engaged when talking to the robot. We, therefore, consider "talking about a movie" as a suitable main task in our setting. However, some participants mentioned in the interview the task as stressful, even though they were explicitly told that the interaction was in no way a test.

However, was the simulation of an autonomous robot listening to them convincing for the participants? In accordance with the WoZ reporting guidelines [67], we report that participants mentioned that they found the robot behavior very sophisticated, or rather basic, as well as very entertaining, or boring. Some of them asked about the natural language processing capabilities of the robot before filling out the questionnaire. In this case, the experimenter asked them to wait with the question until after filling out the questionnaire. The degree of awareness about which parts of intelligent behavior are difficult to achieve on a technical basis, might have influenced their perception.

However, there was no floor or ceiling effect in attitudinal measures, which indicates that the different parts of the behavior implementation (greeting and farewell procedure, idle behavior, and gaze behavior) were adequate.

We did not attempt to categorize the movie genres of the described movies. Arguably, the chosen genre might have had an emotional priming effect. Remembering the plot of a comedy in contrast to a drama might change the participants' perception of the robot during the interaction.

Next, we talk about the limitations of the statistical evaluation. We wanted to

establish a functional relationship between GAR in the range of 0.1 and 0.9. However, it is difficult to produce sample points for the continuous range of 0.1 to 0.9. Therefore we settled on 0.2 increments between conditions.

As we have observed, human gaze changes during the interaction. An avenue of gathering more data across multiple GAR parameters would be to change the GAR of the robot continually during the interaction and interpret the associated gaze responses of the participants. This method might be able to produce detailed insights into continuous ranges of GAR.

As we did explore our hypotheses without prior assumptions, comparing all groups to each other could produce arbitrary relationships, as opposed to a linear regression. However, post-hoc group comparisons for even five groups lead to a very strict alpha level correction. Thus, further increasing the granularity of the independent variable seems infeasible for the empirical topic of our work.

Another difficulty of the granularity of data was the interpretation of gaze sequences. Already with five conditions and three ROIs the resulting table of statistical test results is only useful for researchers, if a pattern can be extracted. However, this allowed different groupings of ROIs to answer different questions (i.e., split of *gaze averted* or *top/bottom*. Additionally, the granularity of gaze sequences is adequate as a basis for learning robot gaze behavior from human data.

Other factors concern the ecological validity of the interaction since the study took place in a controlled laboratory setting, and the relatively homogenous participant sample, namely persons spending time in the TU Wien library. Populations with different age ranges and cultural backgrounds might give different ratings and show different gaze behavior than the participants of this study.

2.5.3 Design Recommendations

In this study, we wanted to determine whether human-inspired gaze settings are indeed the only way to provide agreeable HRI robot behavior parameters. We found, that on a conscious level, participants did not distinguish between the different GAR settings, but there were significant behavioral differences.

Designers who aim for long robot interactions with users should aim for a higher GAR setting (i.e., avert the gaze more often).

Designers should be aware that users will start to imitate the gaze behavior of the robot, with respect to the overall amount of gazing at different ROIs, which might impact interaction goals.

For robot platforms adhering to the honest anthropomorphism principle [20], this means that it should be unproblematic to incorporate additional gaze behavior during a conversation with a specific user. For example, in an elderly care home, the robot could have a conversation with one resident, while performing fall detection monitoring on its surrounding.

We observed that although the robot gaze behavior was not inspired by human parameters, we observed that in some aspects participants reacted as if the robot was indeed a social agent. Thus, sticking to human-inspired values might be a good starting point for a social behavior implementation. However, when technological

aspects demand a deviation from these settings, this does not necessarily lead to a worse perception of the robot by the users.

2.6 Conclusion and Future Work

We conducted an experiment in a controlled laboratory setting in which a robot asked participants to talk about their favorite films. While listening to the participants, the robot used different gaze aversion ratios between condition groups from 0.1 (gazing away only 10% of the time) to 0.9 (averting the gaze to the side 90% of the time). We measured how this factor influenced the user experience through eye-tracking, behavioral measures of interaction duration and word count, as well as self-reported attitudinal data. By varying the independent GAR parameter over a broad range of values that were intentionally not human-inspired, we observed that human participants adjusted their own GAR to the robotic GAR, but not by mere gaze following. Participants showed gaze aversion behavior that is usually associated with uncomfortable interactions when the robot also averted the gaze most of the time. However, the freely chosen interaction duration was the shortest when the robot stared (GAR 0.1) at the participants. Since interaction duration has previously been used as a proxy for comfort in a social interaction, this is not an intuitive finding and suggests that adapting the GAR in an interaction is a stronger mechanism than gaze aversion to regulate the level of intimacy in an uncomfortable interaction. Regarding the self-reported attitudinal measures, participants did not rate a single GAR setting significantly better or worse than others with respect to the *attention* or *capability* of the robot or the *comfort* in the interaction. However, there are behavioral indicators that the GAR settings on both ends of the parameter range were more uncomfortable than medium parameters. These findings suggest that robot behavior designers can use the robot GAR to influence the interaction duration if necessary, with the caveat that intense robot staring might be uncomfortable. In some situations, robots might have tasks that are concurrent with a social interaction (e.g., fall detection, reception work). This might require the robot to avert its gaze to gather information from its surroundings. Our work suggests that in such cases, the robot designers can quite freely adjust the GAR to suit the parallel task, as it does not affect the conscious user experience. To summarize, humans seem to apply social gaze behavior towards robots, even if the robot gaze behavior is not implemented by relying on predetermined human gaze parameters.

Chapter 3

Gaze Sequences

In the previous chapter, we argued for the deviation from human-inspired gaze behavior timings in conversational settings on the second-to-second level to provide more freedom in the gaze behavior design. We will use this insight in conversational settings for joint action settings, where the gaze is also an important modality. In conversational settings, the social component of attitudes toward the robot is more salient than in joint action settings since, during joint actions, there are also other gaze metrics to optimize, like the signaling of task-related beliefs and intentions.

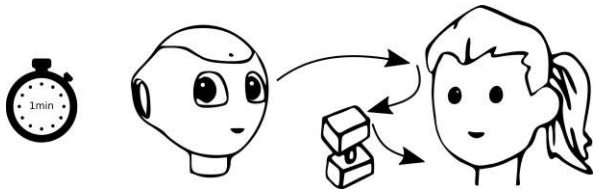

Figure 3.1: In a joint action HRI setting in the time resolution of up to one minute, robotic gaze sequence planning can be learned from empirical human data.

We move downward in the time resolution scale to regard joint attention sequences of about a minute in HRI settings where the human and the robot collaborate on a common goal (Figure 3.1). Within one minute during a collaborative physical task, multiple discrete actions by both actors occur. Both actors must repeatedly choose their gaze target. They switch their gaze between the collaborator's face and the actions they perform, as well as other objects and locations involved in the task. Thus, gaze planning, resulting in a specific gaze target sequence, is essential for joint action tasks. The gaze of an actor can serve as an interaction smoother [86] since it signals the current focus of attention to the collaborator through referential gaze or as a means of coordination through mutual gaze. However, in a joint action HRI setting, it is still being determined how a humanoid robot should dynamically plan the sequence of gaze acts throughout the interaction. Humans solve this by simultaneously and repeatedly using multiple cognitive capabilities, such as sequential planning and ToM.

In this chapter, we review insights into psychological research on joint attention. We derive a structure for a technical implementation of a joint attention gaze controller applicable to human-robot collaboration tasks.

But how can descriptive psychological results be used in robotics? In [21], the authors argue that the width and depth of human coordination capabilities in social contexts will be out of reach for the foreseeable future for technological systems. Instead, they suggest focusing on feasible components that can solve simplified problems or help in a small part of the problem. Heeding this advice, we restrict the problem space. As a task, rearranging colored wooden blocks on a tabletop is an actively used HRI setting [87]. Smooth, cooperative behavior in this setting is challenging both on an engineering and a research level.

Concerning the robot capabilities, the field of automated planning provides a means of finding long horizon high-level plans for pre-specified task domains, such as `Blocksworld`[1] using the Planning Domain Definition Language (PDDL). A prerequisite for this formalism to be applicable in the robotic context is the proper implementation of a state filter and all defined atomic actions. The state filter translates sensor readings to domain-specific states (e.g., whether block A rests on top of block B `on blockA blockB`). Atomic actions are defined in the PDDL domain specification, e.g., the action `pickup blockA` means that the robot can reliably pick up block A without disturbing the rest of the scene. The automated planner finds action sequences from the current state to the specified goal state, in this case, a specific arrangement of blocks on the tabletop. This formalism has previously been extended to collaborative settings [32] in such a way that the robot has a plan indicating who must act and how. Thus, the robot knows when it is its turn to act and when it should wait for the human collaborator to move a block.

However, in previous settings, the gaze modality is only used in a limited way. Mainly, the gaze is statically directed at the tabletop to determine the tabletop state. This method does not yet solve the gaze planning aspect of the scenario.

To this end, we propose adapting the approach in [76], [78] from a static conversational setting to a dynamic joint action setting. This is done by transforming static object roles (e.g., the green block) to dynamic object roles specific to the current plan (e.g., the next object to be picked up at a specific point in time). Using eye gaze data from human dyads, we derive a state-dependent probabilistic gaze controller that produces sequences of human-inspired gaze targets that are suitable for the current state of the tabletop scene.

The resulting gaze controller serves several purposes. First, it produces referential gaze behavior during robot actions. Second, it communicates its belief state by proactively gazing at the objects relevant to the plan execution while waiting for the human to act, with a natural balance of referential and mutual gaze. Third, this serves as a heuristic for improving gaze planning, as imagined in the research area of active vision [33].

We demonstrate the feasibility of the data capture in a real-world scenario and the calculation of the gaze controller in a pilot study. Thus, a humanoid robot has access to gaze planning for the time resolution at the scope of up to a minute.

The following hypothetical situation motivates a joint action scenario which is typical and easy to process for humans. It highlights several aspects of joint attention that are not easily implemented in technological systems. Think of a situation where you have

[1] https://github.com/gerryai/PDDL4J/blob/master/pddl/blockworld/blocksworld.pddl

to coordinate with another person in a physical task at hand. Let us say that you and a friend attempt to move a sofa up a staircase. Both of you have the same goal, namely, to bring the sofa up into another apartment, and the sofa would be too heavy for either one of you to attempt to do so alone. Hence, each of you grabs one end of it. It is also clear to you that your actions influence each other, such that you must monitor and react to each other. Similarly, you can signal to your friend how you imagine squeezing the sofa around the tight corner ahead. You probably will not verbalize each and every intention, but you just push the sofa in one direction more than strictly necessary to signal a direction, or you catch the gaze of your friend by intently looking into their eyes and then gaze into a direction you intend to go. A short nod on their side could signal that they understood. Both of you proceed for a few seconds with the now shared and agreed-upon plan, until you have to check in with your friend to coordinate again.

Collaboration is highly necessary and not overly mentally taxing for humans. Nevertheless, when paying close attention to these collaborative processes that occur almost automatically, it seems that there are numerous different components on different levels of abstraction at work. For example, how do we notice the focused attention of others? Which mental processes let us adapt and align our plans? How do we infer the plans of others? How do we make sure that the other person is really on the same page as us? How do we choose which kind of signal to use for which kind of information? How do we draw the attention of others and signal attention on our part? One must consider all these questions when implementing the capability of human-robot collaboration on a social robot.

In this section, we first contribute a discussion of results in psychology related to this topic (Chapter 3.1). Specifically, we review research on joint attention [88], [10] and ToM [89] with a focus on the human gaze in physical tasks. These are important building blocks generally required for the success of collaborative tasks in human-human interaction (HHI). First, we properly differentiate the two terms and observe how theory of mind builds on joint attention. Then, we focus on joint attention in the robotic context (Chapter 3.2). We contribute a review of different approaches employed by roboticists to provide robots with joint attention capability or at least a technically feasible equivalent. Finally, we propose a novel probabilistic robotic gaze controller for a joint action benchmark between the human and robot proposed by [87], based on building a tower out of various wooden blocks (Chapter 3.3). For object-centered collaborative physical tasks, this represents an approach to generate realistic, intuitive, and interpretable gaze behavior. We report the initial results of a pilot study (Chapter 3.4) and discuss how to include it into the joint action benchmark. Our contribution extends a stochastic gaze controller for static scenarios to dynamic ones. The content of this chapter is based on previously published work in [90].

3.1 Joint Attention in Psychology

Joint attention has been studied since the 1970ies [8]. Research on joint attention in psychology yielded structural and procedural models, as well as analyses whose cues are used to signal the state of joint attention between humans. If we intend to have

service robots in the future that share environments with human beings and provide help in everyday physical tasks, they must be endowed with the ability to engage in joint attention [21] in a similar way as two humans.

Joint attention is the process of sharing one's attention with another person, using social cues for coordination. The coordination effort focuses on a third object, event, or stimulus [91]. One of the earliest reports of joint attention appeared in 1975 in an article by [8] and studied the gaze following ability in infants. The experiment showed that only 30% of two to four-month-old children engage in gaze following, whereas from the age of eleven months, every infant is able to do so. To this day, a significant amount of research is conducted on joint attention in child development.

How can we achieve something functionally similar to human joint attention in *Social Robotics*? First, we consider some results of cognitive and social psychology to better understand how joint attention empowers humans. Furthermore, we consider the components constituting joint attention and how it is embedded in the broader coordination process.

3.1.1 On Theory of Mind and Modeling Joint Attention

One insightful approach is to recognize joint attention as a necessary building block for the more high-level mental capability of ToM. [12] describe joint attention and ToM as relevant in the field of social cognition, as they are concepts explaining how humans process information about other humans in social situations. Children at the end of their second year of life already possess the following capabilities: "1) They understand other persons in terms of their intentions. 2) They understand that others have intentions that may differ from their own. 3) They understand that others have intentions that may not match with the current state of affairs (accidents and unfulfilled intentions)." [12, p. 105]

The term "theory of mind" was coined by [92] and comprises several mental capabilities that develop later in children, around the ages of three to four. It allows them to represent more complex mental states than intentions, namely: "1) They understand other persons in terms of their thoughts and beliefs. 2) They understand that others have thoughts and beliefs that may differ from their own. 3) They understand that others have thoughts and beliefs that may not match with the current state of affairs (false beliefs)."[12, p. 104] [2]

[88], [89] claimed a structural relationship between the separate mental modules of joint attention and ToM. In fact, they claimed that the human ability they call "mind-reading" requires at least four components that build on each other. Mind-reading is defined in the sense that humans can often infer the thoughts, beliefs, plans, and emotional states of other people they observe or think about, in short, reason about "mental things."

[2] Although the term *joint attention* originated in developmental psychology, other approaches in psychology also provided results on the topic, some of which are covered in the following subsections. In these, adults who exhibit a fully developed joint attention capability are the subject of the study. As our robot model is also not developmentally inspired, we do not focus on child development for the remainder of this chapter.

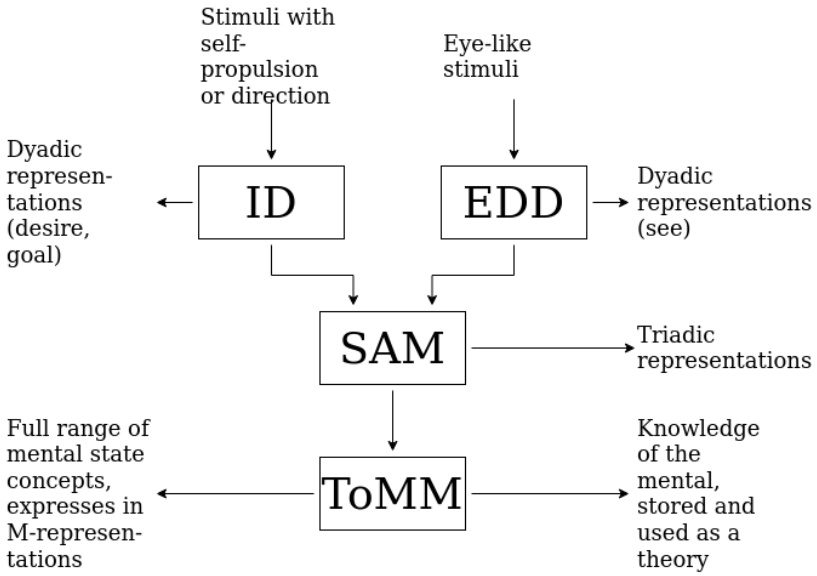

Figure 3.2: Mind-reading system, adapted from [88].

The four component system consists of the intentionality detector (ID), the eye-direction detector (EDD), the shared attention mechanism (SAM), and the theory of mind mechanism (ToMM) (Fig. 3.2). The author claims the modularity to be a necessary part of the model, as different clinical diagnoses can be explained by deficits in specific modules. The ID interprets self-propelled motion of entities in terms of its desires and goals. The EDD specializes in detecting eyes or eye-like stimuli, recognizes the direction of the gaze, and enables the mental attribution of the ability *to see* an observed entity. The purpose of the SAM module is to integrate the two types of information provided by the ID and EDD. This module already allows humans to determine whether another entity has the same target of visual attention. The ToMM module builds on the SAM module and achieves two goals: First, it allows inferring mental states in others from their observable behavior. Second, it allows us to generate explanations for observable behavior by integrating these hidden mental states into theories [93]. ID and EDD form dyadic representations (e.g., a cat chases a mouse (ID), or a cat sees a mouse (EDD)). The SAM module, however, builds triadic representations that are not possible only in the ID and EDD (e.g., I see a cat that chases a mouse). Finally, the ToMM module is able to represent the full range of mental state concepts. These are referred to as *M-Representations* and enable descriptions of mental states where an agent has an attitude toward a proposition (e.g., Johnny believes that "the money is in the biscuit tin."). There is research that builds on this model in the fields of clinical, developmental, and comparative psychology (where the latter studies the mental processes of non-human animals).

3.1.2 Procedural Model of Joint Attention

Another approach to explain joint attention is to categorize processes involved in a successful joint attention event. From observations in infants, the two core processes are

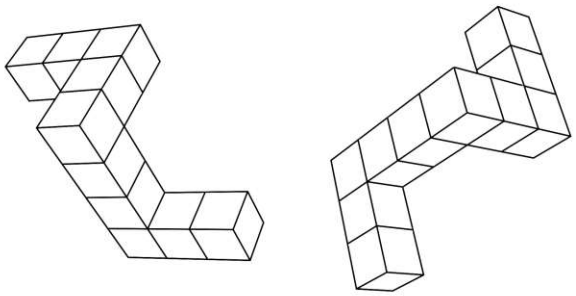

Figure 3.3: An example of a mental rotation task, adapted from [95].

responding to joint attention (RJA) and *initiating joint attention* (IJA) [10]. RJA refers to the ability to follow the direction of the gaze and gestures of others. This allows to establish a common point of reference. IJA describes an infant's ability to use gestures and eye contact to direct the attention of others. Targets of attention are either objects, events, or the infant themself. Clinical research shows that developmental deficits arise in either of these two processes separately. Comparative studies in non-human animals show that animals have the capacity for one of these processes while little to none for the other. Chimpanzees, for example, can respond to, but rarely initiate joint attention [94].

3.1.3 Eye-Mind Hypothesis

The gaze occurs first to gather information, while it also signals information to observers, either intentionally or unintentionally. [95] introduced a simple yet powerful idea, namely the "eye-mind hypothesis." At that point in history, cognitive psychologists strived to understand what was then called the *central processor* of the human mind. Their experiments involved eye-tracking while performing mental rotation of Tetris-block-shaped three-dimensional objects (Fig. 3.3) as well as checking whether displayed sentences correctly described the content of pictures next to them. The authors discovered relations between the ongoing mental operation and the gaze fixation target.

In summary, they found empirical evidence that the "locus of eye fixations reflects what is being internally processed" and that the "locus of the eye fixation can indicate what symbol is currently being processed" [95, p. 53]. The term *symbol* indicates a mental content or entity, something one can think about. For example, when thinking about your favorite mug, your mental representation of that mug is a symbol.

However, there are limits to the eye-mind hypothesis: [96] argue that the eye-mind hypothesis is more likely to hold when a person is performing a visual task, as opposed to pure cognitive tasks or tasks involving modalities other than the gaze.

3.1.4 Types of Gaze Behavior

As discussed in previous sections, there is strong evidence of some connection between the mental focus of attention and the current gaze target. In situations where a potential

interaction partner is present, there are several plausible gaze targets. Looking at objects or specific locations other than the interaction partner is referred to as the *deictic gaze*. When two interaction partners are attending to each other's gaze it is called *mutual gaze*, colloquially eye contact. *Gaze following* is the action of attending to the gaze of the interaction partner, detecting their gaze direction, and then focusing their own gaze onto the stimulus that is being attended by the partner. [97] also disambiguated the state of joint attention from gaze events that appear similar, but have a lower degree of coordination: 1) Simultaneous looking at an object that is triggered by a "pop-out" effect or salient event; 2) Coincidental simultaneous looking at the same object; 3) Gaze following of one agent, while the other pays no attention to the fact that they are being observed; 4) Coordinated gaze at the same object, but attention to different aspects of it (e.g., action intent (like playing with it), or aspect (like color)).

Gaze also plays a large role in pure conversation settings. For example, staring at the other person is often uncomfortable, unnatural, and does not lead to a smooth conversation experience for either participant. Therefore, gaze aversion is often equally important and serves different roles: First, it regulates the intimacy of a conversation. Secondly, it is utilized for turn-taking in a conversation. Gazing at the addressee after an utterance while being silent indicates that the other person should take the floor. Thirdly, averting the gaze indicates cognitive effort. Thus, a speaker can signal that they are not yet done with their turn, even though they are currently silently formulating a statement in their mind [23].

3.2 Joint Attention in Human-Robot Interaction

An envisioned goal for Social Robotics is close collaboration between humans and robots, reaching beyond humans and robots working on different subtasks that lead to a common end result (e.g., pick-and-place robots in production). Actual collaboration between humans and robots is a sequence of shared actions toward a shared goal and requires coordination [98]; in other words, joint attention is employed in the sofa moving example mentioned in the introduction. In our work, we explicitly focus on HRI use cases surrounding object manipulation (e.g., picking up objects) and exclude settings with a stronger social focus.

There is no definitive theoretical model for joint attention on a robot. For implementation purposes, one approach is to view the desirable input-output relation for a given scenario as the requirement and use whichever technique is available and achieves the result. For example, a human and a robot can both generate plans for solving a given problem, but their specific methods can differ.

Additionally, [21] argued that the width and depth of human coordination capabilities in social contexts will be out of reach for technological systems in the foreseeable future (although constant progress is being made). We must instead direct our attention to AI research and look for feasible components that solve simplified problems or help with a small part of the problem.

The authors split the problem of developing a ToM for social robotics into a micro (actual interaction), meso (relationship building), and macro level (roles and persona).

On the micro level, they associate ToM, perspective-taking, shared intentionality, and common ground. Common ground refers to mental content of which all interaction partners know that this content is known by everyone.In relation to these levels, our work addresses a joint attention implementation on the micro level, excluding considerations on the meso and macro level.

3.2.1 Implementing Joint Attention for HRI Tasks

HRI research has produced several results regarding joint attention implementations on robots. These include the capability of drawing attention to another reference point, as well as establishing, monitoring, and ensuring joint attention during an interaction. The interaction settings are either conversational with different points of interest in the environment or physical such as object handovers or other object manipulations.

These scenarios differ from pure conversational settings between a human and a robot. Typically, joint attention HRI settings involve at least another object, location of interest, or event besides the two agents. The human and robot both measurably focus their attention on this third entity or even physically interact with it. [99] proposed an HRI joint attention mechanism. They presented the difficulty of drawing a person's attention to another reference point. This includes how to make a person understand the communicative intention of the robot, and how to deal with the person's attention status. They implemented the pointing and gazing functionality on a humanoid robot, enabled the robot to perform the mutual gaze, and represented the person's attentional focus as a spatial coordinate. They conducted an experiment, where the robot acted as a presenter of a scientific poster to a human participant. Results indicate that humans gazed more frequently towards the poster when the robot acted according to the proposed attention mechanism.

[11], [100] extended the Responding and Initiating Joint Attention (RJA, IJA, Chapter 2.2) model by an explicit Ensuring Joint Attention component (EJA). The EJA component in their framework encapsulates the ability to monitor another's attention to verify that joint attention is reached and maintained. They describe a cannonical joint attention episode between two agents comprising five steps: 1) Connection of two agents, where they become aware of one another and anticipate an interaction; 2) Joint attention request by the initiating agent, where it focuses the attention on a third object and uses communicative channels such as pointing, gesture, and voice; 3) Joint attention response, where the other agent also focuses on the third object; 4) Monitoring, where the initiating agent ensures joint attention by switching the focus between the other agent and the referential focus; 5) Joint attention is reached, the interaction continues. The authors equipped their social robotic platform with a finite state machine, a procedural representation of the described joint attention episode. The perception capabilities of the robot included face detection, marker detection to perceive pointing actions, and speech recognition for a few phrases, which were used to check the attentional state of the human interaction partner. The humanoid robot had a movable head with two degrees of freedom and eyes with two degrees of freedom, as well as movable arms for pointing and a speaker for verbal communication. The authors conducted several experiments. In the first one, the robot had to show that it

can respond to joint attention by attending to objects that the humans pointed at. In the other experiments, which were video-based, the robot had to direct the attention of a human to a presentation as a tour guide, ensure attention while delivering a verbal message and giving directions. The overall result indicates that robots with their joint attention implementation performed better in responding to pointing actions tasks and were considered more natural in the video-based experiments. [11] mentioned that it is unclear how to design the specific timings of the EJA component.

[101] created an autonomous gaze system for the Furhat robot (a mounted mannequin head with an animated video-projected face) for a puzzle-like spatial reasoning task conducted on a tabletop. Their attention system is split into a proactive and a responsive gaze layer with different priority levels. Gaze events of higher priority override those with lower priority. The timing of gaze shifts is uniformly sampled from predefined ranges. The human participant, task objects, and the surrounding environment (for gaze aversion) are possible gaze targets. The proactive layer handles the gaze related to the speech acts of the robot (eye contact, IJA at task objects) and idle gaze behavior through gaze aversion. In the responsive layer, user speech activity and a detected mutual gaze led to a mutual gaze, while gaze tracking and object tracking was used for RJA events to gaze at objects. The system was then used to engage with the user during the task, comment on their progress and provide hints for the correct move. In a user study, self-reported data suggested that the robot with both responsive and the proactive layers was perceived as more socially present than the robot with only the proactive component, as only the former was able to react to the user and thus engage in joint attention.

Joint attention capabilities have also been shown to improve collaborative physical tasks like handovers in HHI [102], but also HRI. [103] created a two-layer architecture for physical robot-to-human handover tasks for a humanoid robot. The first layer represents the physical state of the handover as a Hidden Markov Model with the states “Robot pick up,” “Robot hold,” “User grap,” and “Robot not hold.” These states, however, are only estimated by the current and torque values measured in the robot hand. A higher-level layer was then added that serves as an additional safety check to release a grasped cup to the human under the right conditions. The authors observed that human users performed a sequence of actions in a successful handover: browsing the environment, looking at the target cup (optionally looking at the cup repeatedly), and finally grasping the cup. The second layer registers the gaze pattern of the human by monitoring the head direction. Only if the described gaze pattern is detected before registering a grasp attempt the robot releases the cup. The extension of the handover architecture has been empirically shown to result in fewer unsuccessful grasp attempts.

Similarly, [104] compared HRI handover scenarios with varied humanoid robot gaze behavior. In an HHI handover study, they detected two gaze patterns of the agent handing over the object: The shared attention gaze is gaze-directed at the projected handover location. In addition to this behavior, a turn-taking gaze pattern sometimes occurs, which consists of establishing eye contact while reaching out. These findings were implemented in a humanoid robot, which resulted in the experimental conditions of no gaze (baseline), shared attention gaze, and the shared attention gaze plus turn-taking cue. The authors found that human users reached for the handover object earlier in the

two gaze conditions and reported a trend of self-reported preference for the turn-taking behavior over the other two conditions.

3.2.2 Planning for Joint Human-Robot Interaction

As [88] mentioned, humans are expert mind readers. Hence, when a human observes another human in an everyday situation, the observer most likely forms an idea about what the observed person is trying to achieve with their current actions. For example, if you see someone in a kitchen opening the cupboard drawer containing all the mugs, you will probably already think about which drink they want to consume, while all they did was simply open a drawer. Notable, it is quite possible that the observed person will do something different, but our experience tells us that getting a drink is the most probable goal given such an observation. One research direction on Joint HRI is to explore methods for simulating this human capability, namely AI planning.

We distinguish between symbolic and subsymbolic planning: In a formal language, symbols are atomic tokens of a language. This means they cannot be split into smaller units of meaning. Symbols are manipulated with some kind of procedure to build more complex expressions. This is (mostly) comparable to our spoken language with its single tokens, such as "cat," "in," and "tree." From these tokens one can build expressions "cat in tree" or "tree in cat." One of these makes more sense from our experience than the other, but both are correct expressions in our language. In turn, the expression "cat tree in" would not be considered part of our language. There is simply no valid symbol manipulation sequence that can generate this expression. Nevertheless, symbols alone do not have any meaning in themselves, and the problem of assigning symbols to references in the physical or social space is referred to as the *symbol grounding problem* [105], [106]. In contrast, subsymbolic planning involves a more direct representation of the problem. Consider a map where one must find the shortest route between two points. There are no tokens that are manipulated, just path-finding reasoning with the data provided by the map.

Generally, subsymbolic planning is often used for collaborative problems such as social navigation (i.e., safely moving through a crowd of people [107]) or human-robot handovers, where the problem is represented and solved in a task space like the Euclidean space of a suitable dimension. For more abstract or high-level planning problems, however, a symbolic approach makes the problem formulation more compact. In this book chapter, we focus on such representations.

Before formulating the problem itself, however, we must consider our underlying assumption, namely the rationality of all involved agents. Broadly, this means that an agent would rather perform an action that results in a benefit to them rather than harm. In the frame of the problem definition, the question is how to define a cost function, or even how to know that optimizing the *expected* cost for a problem is even the right thing to do [108]. Assigning reward (or cost) values to certain outcomes of a decision process may be intuitive. These may be of a monetary value, or of a more subjective value, like choosing between washing the dishes or sweeping the floor. Thus, every action is assigned a reward value. If the action outcome is stochastic, then a reward distribution is assigned to each action. An example of this is a game where an agent

chooses between receiving 1000 € or letting a coin flip decide whether they receive 2000 € or nothing. Although the expected value of both actions is the same, most people will have a preference for one or the other, depending on their inclination toward gambling. Thus, using the expected value alone is insufficient to model the preferences of agents. This is solved by deriving a so-called utility function for all action outcome distributions. For a utility function to exist, a rational agent must be able to provide a consistent ranking of different probability distributions over outcomes according to the axioms of rationality [108]. Thus, each action outcome is assigned a utility value. Finally, a cost function can be derived from the utility function.

Markov Decision Processes (MDP) can be used to solve problems in sequential decision theory [108], where agents repeatedly chose actions according to their current state. A single agent MDP is defined by 1) a non-empty *state space* S, which is a finite or countably infinite set of states; 2) for each $s \in S$ a *finite, non-empty action space* $U(s)$ with a *termination action* (it is applied when reaching a goal state); 3) a finite, non-empty *nature action space* $\Theta(s,u)$ for each $s \in S$ and $u \in U(s)$ (a *nature* decision maker represents uncertainty in the action outcome); 4) a state transition function f that produces a state, $f(s,u,\theta)$, for every $s \in S$, $u \in U$, and $\theta \in \Theta(s,u)$; 5) a set of *stages*, which is either infinite or set to a fixed, maximum stage (i.e., how many sequential actions can be taken before the problem must be solved); 6) an initial state $s_I \in S$; 7) a goal set $S_G \subset S$, and 8) a stage-additive cost functional L. The goal of the agent is to find a plan to reach a goal state from the initial state. Because there are stochastic state transitions, a policy $\pi : S \to U$ must be found for all $s \in S$ that minimizes the cost. Alternatively, π can be a mapping from a state to a probability distribution over the action space. Then, this corresponds to a *randomized* instead of a *deterministic* strategy.

Markov chains are a simplification of this model without an explicit decision maker. Nature determines the outcome of the next state alone. Markov chains are used to model stochastic processes and, like MDPs, fulfill the Markov assumption (Equation 2.1). $S_1, S_2, \ldots, S_t$ denotes the sequence of random variables up to timestep t, where the outcomes are $s_i \in S$. This means that only local information, and not the entire history of the process is used to determine the probability of the next state transition.

Generally, artificial agents have some sensing capability to determine the current state they are in. However, due to nature, sensor errors can occur. This leads to another type of uncertainty besides stochastic state transitions, namely state uncertainty. This means that the agent does not know for sure whether it is in a single current state $s_t \in S$, but holds a *belief* about the current state, expressed as a probability over S. Including this belief into planning lifts the problem formulations from the state space into the state belief space.[3]

For joint action scenarios, it is important to model more than one active decision maker. This leads to the inclusion of the game-theoretic concept of the *two-player nonzero-sum game* [108]. One formulation is to extend the MDP definition by another agent. Herein the two agents (players) P_1 and P_2 have their respective action spaces U_1 and U_2. In zero-sum games, there is only one cost function $L : U \times V \to \mathbb{R} \cup \infty$, which

[3]Literature presented in this chapter as well as our contribution only concerns planning in state space.

one player regards as reward, and the other player as cost. In the nonzero-sum game, however, each player has a different cost function (like L), namely L_1 and L_2. Both players now aim to minimize their costs according to their respective cost functions. Thus, in such games different degrees of cooperation can be formulated, from total cooperation to a zero-sum game. This formulation can be lifted to sequential games on game states by expanding the MDP definition by another player.

In symbolic planning problems, if the planning problem uses deterministic action outcomes, a widespread approach in robotics is to employ *classical planning*. A *classical planning domain* (i.e., a *state-transition system*) is a triple $\Sigma = (S,A,\gamma)$ or a 4-tuple $\Sigma = (S,A,\gamma,cost)$. S is a finite set of possible *states* of a system. A is a finite set of *actions* that an actor can perform. $\gamma : S \times A \to S$ is a partial function called the *state-transition function*. When $\gamma(s,a), s \in S, a \in A$, is defined, then a is *applicable* in s, and $\gamma(s,a) \in S$ is the outcome of the action. $cost : S \times A \to [0,\infty)$ is a partial function with the same domain as γ, defining a metric, which is to be minimized, such as the monetary cost or time. In this kind of representation, there are the assumptions of a *finite, static environment*, *no explicit time* (except the cost, if it is to be interpreted in this way), and *no concurrency*, indicating that actions cannot be performed in parallel. Actions are *deterministic*, which means that the outcome of an action is known with certainty [109].

In the formulation above, there is a finite set of states ($S = (s_0, s_1, \dots)$) with no specific relation to one another. A more succinct way of describing states is by using *state-variables* (*predicates*) and *objects*. Hereby, states are defined as specific instantiations of these state-variables. These state-variables can use objects as arguments. A concrete example is the PDDL planning domain `blocksworld` [110], which is a formal planning language that is commonly used for robotic tasks that involve planning in semantic domains. It is an approach to encode a classical planning problem derived from previous formal languages like the *Stanford Research Institute Problem Solver* (STRIPS) [111]. A PDDL problem is encoded by a domain and a problem instance, where the domain describes the state-variables and operators, which are uninstantiated action templates. Once an operator is given parameters, it is called an action. Operators, like `pickup`, are defined with objects as possible parameters (`?ob`), preconditions, and effects. Only when the preconditions are met in the current state, the action is performed by applying the effects of the action on it. This is done by adding and/or removing predicates from a state. The problem instance describes the existing objects, the initial state, and the goal. The solution represents a plan which solves the problem. There are PDDL versions that allow durative and concurrent actions, continuous and conditional effects, etc. However, we disregard these options for simplicity.

3.2.3 Plan Recognition in Classical Planning

Classical, symbolic AI planning is an approach to endow a robot with a planning capability suitable for joint HRI situations. However, it is only a part of the solution. A robot must also be able to infer the goal and plan of the interaction partner. To this end, classical planning plan recognition is employed [26], [112]–[114]. An advantage of this approach is the reuse of the planner that the robot uses to generate its own

plans. The plan recognition problem is formulated as a triple $T = \langle P, G, O \rangle$, where P is a planning domain, G is a set of goals, and O is a sequence of observed actions. See Chapter 4.2.2 for a formal definition. When the sequence O ends in a state that is a goal, the goal recognition is trivial; however, when the observation ends in a state that is not a goal, the problem is to predict which is the most likely goal, to rank these goals with regard to their relative probabilities, or to assign probabilities to the different goals. Various approaches have different ways of executing this, but their commonality is to transform the original planning domain to accommodate the observations and subsequently compare the cost of different plans. Different plans are generated for a single goal, e.g., one that satisfies the observations and one that does not. When the cost of adhering to the observation for a goal is significantly higher than reaching the goal without doing so, that goal is probably not likely to be the actual goal of the observed actor. This builds on the assumption of *rationality* of an agent, i.e., that one attempts to fulfill their desires in an effective and efficient way.

3.2.4 A Benchmark in HRI for Joint Action

Situations that are simple and intuitive to solve for a human team, such as building a specific tower out of wooden blocks on a table, prove to be complex and difficult for current joint attention research. Therefore, this setting - a human and a humanoid robot who attempt to build a block tower - is used as a recurring scenario in joint action research [7], [115]–[117].

Pure plan recognition research often only treats problems that are already formulated in formalisms like PDDL. Similarly, the problem formulation of plan recognition does not deal with the continuous coordination effort that is necessary in joint attention situations. [32] combined classical planning in the block world domain with the demands of joint action problems. In their study, they set up a joint action scenario with a human participant and a PR2 robot[4] (Fig. 3.6, left). The PR2 robotic platform was equipped with several optical sensors and two arms with pincer grippers. The setup includes fiducial markers on the blocks to facilitate their recognition. The robot was able to perceive the world state (i.e., the current arrangement of blocks) and manipulate the blocks.

The robot and the human participant have a shared goal. They stand on opposite sides of a table and attempt to build a specific block tower with blocks lying on the surface. However, each agent is only able to reach some of the blocks, hence they must collaborate. To introduce another challenge, there is not one single fixed sequence that results in the correct block tower (Fig. 3.4). For example, there are two places for putting the red blocks and each actor has access to one of the two red blocks. They need to coordinate who picks which placement spot. The following difficulty arises when the agents must place the block stack green-blue-green. Again, each actor has access to only one green block. Thus, the actors must coordinate who places the first green block.

The authors approach this scenario as a multi-agent planning problem. The robot

[4] https://robots.ieee.org/robots/pr2/, Image source: https://www.wevolver.com/wevolver.staff/pr2

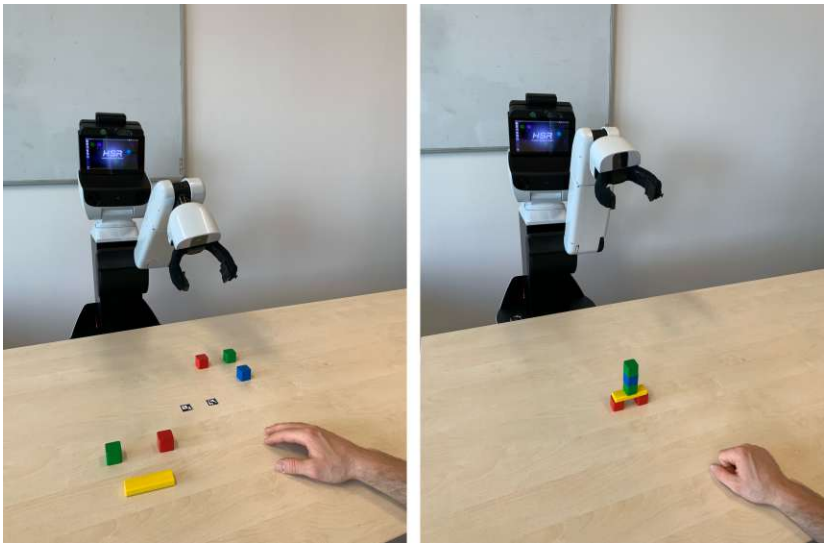

Figure 3.4: Joint action task described in [32]. Left: Initial configuration. Right: Goal State.

finds plans by modeling three discrete actors (itself, the human, and a fictitious *X agent*) who can place the blocks. In valid plans, actions that are assigned to the *X agent* mean that either of the two actorsm, human or robot, will perform the action. Notably, in the example above, there could be multiple open actions at once, e.g., placing the two initial red blocks in the center. In the shared plan, when the next necessary step is an action performed by the human, the robot waits for its completion. When the next necessary step is a robot action, the robot performs it. However, whenever an action is assigned to the *X agent*, the robot has different approaches for enacting this shared plan, namely acting lazily (i.e., waiting for a specified amount of time and watching whether the human will perform the action) or in a hurried way (i.e., the robot always attempts to immediately perform an *X action*. Furthermore, agent assignments can change during the plan execution, such that the plan must be recalculated after each step. For example, when one actor places the first green cube, the placement of the second cube is no longer an *X agent* action, as only the other agent has a green block left. This demonstrates the complexity of this simple collaborative block world problem as it already exposes numerous interesting and difficult aspects of joint action and requires further research effort. Thus, to establish a standardized scenario, [87] propose a joint action scenario similar to [32]. Their goal was to facilitate finding answers to the following questions: "What knowledge does a robot need to have about the human it interacts with [...]?"; "What information should the human possess to understand what the robot is doing and how the robot should make this information available [...]?" [87, p. 2] The proposed simple HRI scenario has the following setup and assumptions:

The common goal of the human and robot is to build a stack of four blocks in a specified order with a pyramid on top. They are on opposite sides of the table and face each other. Each agent has access to two of the four blocks. There are two pyramid pieces, one on either agent's side of the table. Only one of the two agents is supposed to place the pyramid piece at the end of the action sequence. The agents are restricted to

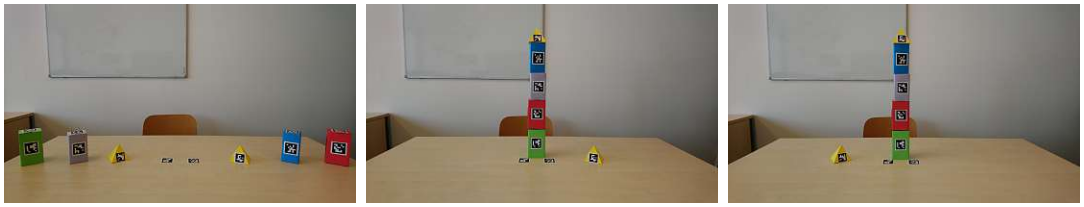

Figure 3.5: Left: Initial configuration. Middle and Right: The two possible goal states.

Figure 3.6: Two domestic service robots. Left: Toyota Human Support Robot (HSR). Right: PR2 by WillowGarage.

the actions of the block world domain, plus a handover action, and a possibly support tower action.

Fig. 3.5 illustrates the initial and the possible goal states. Both agents are assumed to perceive the current world state and thus are able to locate objects and assess their reachability by either agent. Finally, each agent is able to observe actions of the other.

3.3 Toward a Gaze Mechanism for Joint Actions

As described above, one of the two core questions posed by [87] is *how a robot should signal information that is important to the human in order to enable smooth collaboration.* We argue that the gaze is a useful modality for this specific benchmark task even for robots, as it is highly intuitive for humans to interpret, and is perceived constantly without being bothersome (in contrast to continuously verbalizing information, for example). It is furthermore potentially easier to perform than other non-verbal behavior, e.g., pointing.

Conveniently, common mobile service robotic platforms such as the PR2 by WillowGarage or the Toyota Human Support Robot[5] (HSR) (Fig. 3.6) have head-like extensions with two degrees of freedom that house forward-facing optical sensors. Therefore, the head orientation represents in fact the direction of gaze. Social humanoid robotic platforms, such as Pepper from Softbank Robotics[6] or Nao[7] (Fig. 3.7) have the

[5]https://robots.ieee.org/robots/hsr/, Image source: https://developer.nvidia.com/embedded/community/reference-platforms/toyota-hsr

[6]https://www.softbankrobotics.com/emea/en/pepper

[7]https://www.softbankrobotics.com/emea/en/nao

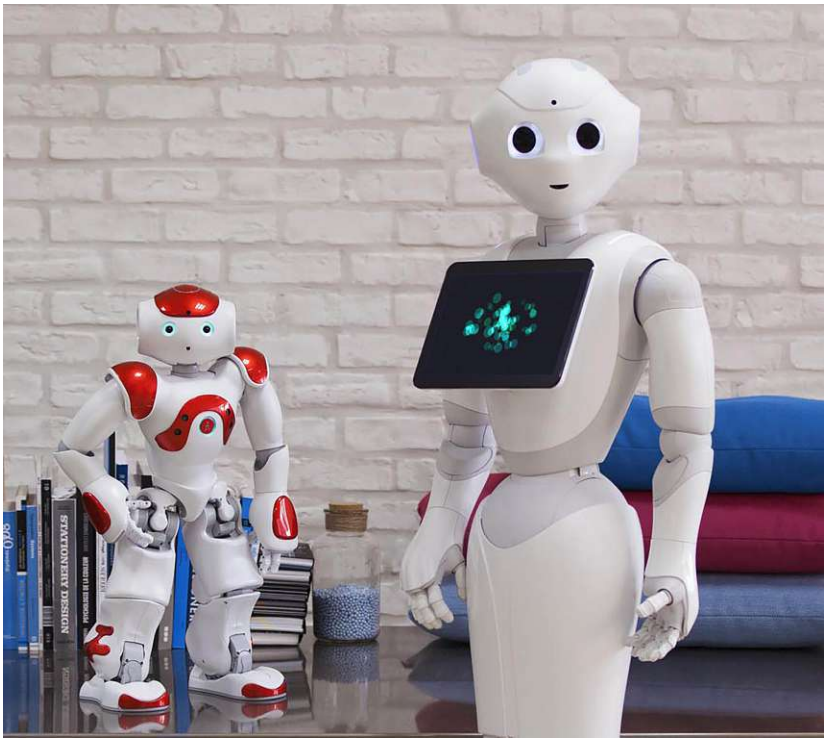

Figure 3.7: Two social humanoid robots by Softbank Robotics. Left: Nao. Right: Pepper.

same degrees of freedom in their heads and have already been used in gaze-related HRI studies. Research has shown that their head orientation communicates attention [118], [119] and is interpreted as gaze by human participants. We, therefore, propose that the gaze in the joint action benchmark will significantly smooth the interaction between the human and the robot, as it has previously in the different communicative HRI settings surveyed by [22].

3.3.1 Comparison of Human-Derived Gaze Mechanisms

It is important to model the gaze behavior of domestic service robots in a way that it primarily does not impede their functionality, and secondly serves a communicative purpose in joint attention and joint action situations. The human gaze is very effective at doing both simultaneously. During object manipulation tasks, humans gaze at task-relevant objects and locations [120], [121]. This behavior is a rich source of information for an interaction partner in collaborative scenarios. In the ideal case, a robot would use its gaze to improve its belief about the current world state, as well as utilize the communicative aspect of gaze. Therefore, a model of the human gaze in joint action tasks can be used as an initial heuristic. The most important characteristics of such a model are the gaze locations and timings, i.e., when to look at what. Another, perhaps less important factor, is the transition dynamics, i.e., which animation profile is exhibited by gaze transitions.

When implementing a gaze model for a robot that interacts with another actor and

objects in its environment during a joint task, the question of *when the robot looks at a specific gaze target needs to be addressed*. More specifically, which sequence of gaze targets and fixation durations communicates the attentional (gaze) focus of the robot to the human actor? We assume that the gaze is divided between the objects the robot manipulates itself, the object manipulations of the human partner, and the human's hands and face. The gaze at the objects that the robot wants to manipulate is (at least at some point in the process) necessary for the proper execution of the planned action. Thereby, the robot communicates its own attentional focus through gaze. The gaze at the object manipulations by the human is necessary to assess the current world state. The gaze at the face of the human is necessary to ensure the joint attention status. Similarly, at each point, the gaze of the robot could be interpreted by the human to draw conclusions about the attentional state of the robot.

This might seem to overly complicate the block stacking benchmark task, however, it represents only an initial step to solve more difficult scenarios. Examples of these include tasks with more than two actors, and tasks that include more movement, such that not each important location of attention is captured in a single camera angle, for example when objects are positioned further apart, when actors do not face each other all the time, or when objects are occluded.

3.3.2 Modeling the Sequence of Gaze Targets

Next, we discuss how to create a gaze model for the above-mentioned tasks. [77], [78] employed a specific methodology for creating a gaze controller specifically for gaze aversion in conversational settings. They recorded two eye-tracking datasets in dialogs between two humans, where one participant was the interviewer and the other the interviewee. One dataset was generated from the view of the interviewer, the other one from the view of the interviewee, using a wearable Tobii Glasses 2[8] eye-tracker. For each interview perspective they used a sequential data mining method to derive the most common gaze shifts, where the following gaze targets were encoded: the face of the dialog partner (referred to as *gaze contact fixation* by the authors), and gaze aversion directions relative to the position of the face (down, up, left, right, and diagonal).

More importantly for this book chapter, stochastic models are also used to model gaze sequences. (First order) Discrete-Time Markov Chains (DTMC) describe sequences of gaze directions using the Markov property assumption (Equation 2.1, Section 3.2), i.e., only the previous gaze target determines the probability of the next gaze direction and the possible states are in the set $\Omega = \{center, up, down, left,$
$right, up-left, up-right, down-left, down-right\}$.

A simplifying assumption was made, namely time-invariance, meaning that the probabilities do not change depending on the position in the sequence. This allows the gaze model to be represented as a Markov chain transition matrix of size $|\Omega| \times |\Omega|$. A cell matrix cell value p_{ij} represents the probability of changing the gaze from target x_i to x_j and the rows must sum up to 1.

The authors argued that a gaze controller producing such stochastic behavior will

[8] https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

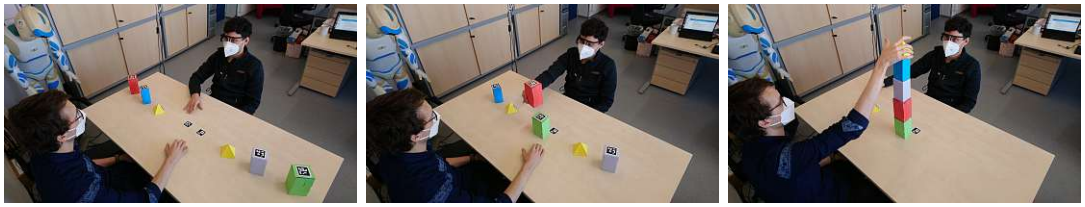

Figure 3.8: Gaze data capturing during the pilot study. Left: Initial position. Middle: Eye-tracked participant places a block from the reachable area. Right: Placement of the pyramid block. Both participants can place their pyramid, and after a negotiation phase, the other participant places the final piece.

be helpful in HRI conversational settings. Further, they have future plans to validate this idea by implementing it on a humanoid robot and conducting HRI validation studies following the methodology of [23], where the proposed model with proper gaze timings was tested against a baseline with static gaze and a baseline with inverted timings ("anti-timings"). The study argued that both baselines should lead to a worse evaluation of the robot by the human interview partners than the proposed model.

This kind of gaze control is aimed at conversational HRI settings and has numerous useful applications, such as tour and info guidance, receptionist duties, etc. Mobile service robots such as the Toyota HSR can additionally perform object manipulation tasks and require gaze control for them, as argued above. Providing a gaze controller for the joint action benchmark task described earlier is thus helpful to handle more realistic scenarios in the future.

3.4 Data Collection

We describe how to adapt the procedures from [77], [78] to a collaborative object manipulation task. In a pilot study, we recreated the block stacking task with the pyramid top presented in [87] (Fig. 3.8). Two human participants sit opposite each other at a table. One of the two participants per trial wore a PupilLabs Core[9] [122] eye-tracker with monocular eye-tracking.

We tested two pairs of participants ($n = 4$). Each pair conducted two trials. After the first trial, they swapped positions, such that each participant wore the eye-tracker in one trial. All participants were briefed by the experimenter. The participants were asked to read and sign an informed consent form. They were instructed to collaboratively build a specified tower (from bottom to top: green - red - lavender - blue - pyramid). Figure 3.5 depicts the view of the person wearing the eye-tracker. This person was instructed to act as if only the red block, blue block, and right pyramid is reachable for them. The person sitting opposite was instructed to act as if they can only reach the green block, the lavender block, and the left pyramid.

The participants were instructed to follow a set of rules: (1) Use only your right hand. The task was simple enough for humans, such that non-disabled persons can use their right hand even if it is not their dominant hand. (2) The right hand is supposed to

[9] https://pupil-labs.com/products/core/

always be above the table. (3) The left hand is supposed to be out of sight underneath the table. (4) Participants were asked not to rotate the blocks while moving them.

The participants were informed that this was not a test and that speedy execution is not important. Starting a grasping action while the other person is still placing their block was not forbidden. The blocks display fiducial markers facing the person wearing the eye-tracker and participants were asked to grasp the block in a way that does not occlude the markers. The placement position of the bottom block was also marked on the table with fiducial markers. These rules and restrictions were implemented such that the resulting behavior is similar to the one of a robot during such a task.

The two participants were asked to memorize and recite the correct block stacking sequence before the experiment to avoid execution mistakes and to limit gaze and other behavior that is not associated with shared plan execution. The participants were not allowed to discuss any strategy before the task and were not allowed to speak during its execution.

The participant wearing the eye-tracker is referred to as the *robot* (R), because the recorded gaze behavior is meant to be implemented on a service robot. The other participant is referred to as *human* (H). X denotes the *X Agent* (X). The resulting interactions included only actions that were in accordance with the optimal plan:

```
(pickup H green) (place H green table) (pickup R red)
(stack R red green) (pickup H lavender) (stack H lavender red)
(pickup R blue) (stack R blue lavender) (pickup X yellow)
(stack X yellow blue)
```

Gaze behavior that results from these interactions thus depicts gaze behavior for smooth interaction without errors. During the last step, where the two agents need to negotiate who picks up their pyramid piece, gaze behavior indicative of negotiation will take place. The generalization is naturally only possible for an appropriately large sample size and only for populations with the same demographic properties. In this chapter, only a preliminary feasibility check with a small sample size is presented, and the obtained results serve as an exemplary outcome.

The goal of this experimental setup is to elicit successful collaboration and the corresponding gaze behavior in the person wearing the eye-tracker. Large-scale plan re-negotiations during the task must be avoided. Small-scale negotiations (i.e., resolution of *X agent* actions) fall within the capabilities of the planning formalism. This choice is motivated by the consideration of the full robot architecture: In problems that are more general than the chosen experimental setting, large-scale plan deviations might occur. However, after each action (planned or unforeseen), the visual sensors of the robot will detect the resulting world state, which will be used as the initial state of the planning problem. Then, a new shared plan will be calculated. This might result in a new planned sequence of actions. The robot gaze controller always acts with respect to a determined plan, as described below in further detail. Thus, if a new plan is calculated, the gaze is adjusted according to the newfound plan. Plan changes occur due to unforeseen actions; however, this does not result in unspecified gaze behavior. The robot gaze always corresponds to the belief of the robot and visualizing the belief of the robot through gaze is the goal of this gaze controller.

During the trials, the strategy to overcome the ambiguity of who places the pyramid

	Next Target						
Target	Face	Hand	Table	Green	Red	Lavender	Blue	Yellow
Face	0.12	0.12	0.29	0.17		0.17		0.13
Hand	0.13	0.23	0.02	0.22	0.11	0.11	0.07	0.11
Table	0.11	0.37	0.08	0.25	0.04	0.04		0.11
Green		0.30	0.05	0.24	0.14	0.05	0.17	0.05
Red	0.10	0.10	0.25	0.12	0.23	0.10	0.10	
Lavender	0.38	0.07			0.07	0.11	0.26	0.11
Blue		0.19	0.04	0.11	0.04	0.14	0.48	
Yellow	0.67	0.17	0.08	0.08				

Table 3.1: DTMC transition probabilities of eye-tracked locations.

was always solved with the "turn-taking" strategy, where the person who placed the topmost rectangular block waits for the other person to place the pyramid. In our small sample, the placement of the pyramid occurred either immediately or after a short period of inactivity.

For each gaze data sample, we conducted the following evaluation: Using fiducial markers[10], as well as (the partner's) hand and face tracking [123] allowed the recognition of these objects in the eye-tracked video. By defining a 100 pixel radius around each target, we distinguish eye fixations of the other person's hand and face, as well as the placement location of the bottom block on the table, as well as all other blocks and pyramids. Furthermore, we encode fixations gazing at none of the above.

For each sample, a sequence of fixations is extracted from the gaze data, and we create a DTMC transition model by counting the transitions. In this scenario, this yields a 8×8 matrix (pyramids are counted as one object). The gaze targets are the face of the partner, the hand of the partner, the placement location on the table, the four blocks, and the two pyramids, which are counted as one object due to their interchangeability.

For this gaze controller, we disregard fixations that do not fall in the radius of any target. If a fixation falls on a spot in the visual field that is currently in the radius of more than one target, we count split transitions and mark more than one object as currently active until the gaze falls on a single object again.

The aggregated model in Table 3.1 was derived with the gaze model for every sample. There are two possibilities of arriving at the probability values, which sum up to 1 per row: Either the frequency counts of the transitions are averaged per sample, and then the averaged matrices are added and again normalized per row. This is the variant we chose since it leads to an equal representation of each sample. Another method is to add all frequency count tables and only then normalize over the rows.

The controller can then be applied to create gaze behavior by choosing a basic timestep unit, e.g., one second (This varies with the task, and the robot embodiment.) and creating a gaze sequence by starting in a random or predetermined (e.g., face) state. The next state is sampled with the probability weights of the row of the current state.

Further work is planned to split the gaze controller into two parts and to analyse

[10] https://april.eecs.umich.edu/software/apriltag

				Next	Target		
Target	Face	Hand	Past	Prev.	Curr.	Next	Future
Face	0.08	0.08	0.19		0.11	0.19	0.33
Hand	0.16	0.19		0.09	0.27	0.20	0.09
Past	0.11	0.11				0.78	
Previous			0.12	0.12	0.12	0.12	0.50
Current	0.20	0.35			0.19	0.22	0.03
Next	0.23	0.12	0.11	0.06	0.31	0.15	0.03
Future		0.33			0.50	0.17	

Table 3.2: DTMC transition probabilities of eye-tracked locations in their dynamic context of the plan execution.

whether the gaze behavior in the action phase (placement up to the last block) differs from in the negotiation phase (placement of either pyramid).

3.4.1 Creating a Gaze Controller for Time-Variant Scenarios

Table 3.1 indicates the specific objects the participants gazed at during the whole task duration. This neglects an important factor, namely the dynamic nature of the time-variant task. During the task, the world state is defined by the block arrangement and whether an actor is currently grasping a block. It is clear to both actors which block to grasp next (or whether to negotiate who should place the pyramid top). For the plan execution, the following block to be placed has another role to the actors of the current action than a block that has already been placed. Therefore, we annotate the video samples with the current state of the world, i.e., which blocks have already been stacked (neglecting whether a block is grasped or not). Thereby, we partition the set of blocks, pyramids and table placement location into sets of *past*, *previous*, *current*, *next*, and *future*. The *current* block is the one that must be picked up and placed at a specific point in time. The *previous* block is the block that was placed right before the current block. Prior to placing the first block, *previous* indicates the table placement location. The *next* block indicates the block to be placed after the current block. *Past* and *future* blocks group blocks that have been placed before *previous*, and must be placed after *next*, respectively. The controller in Table 3.2 is derived with this dynamic assignment of object roles. Hence, we preserve the time-invariance assumption of the gaze controller with this transformation from block identities to temporal roles.

3.4.2 Future Work

We tested the described pipeline to derive a gaze controller with transition probabilities based on a larger sample size. Careful attention to the validity of the result must be paid, as numerous design choices have been taken in the aggregation method of the different study participants and filtering of fixations in single samples. Therefore, we propose a validation study where a pre-programmed humanoid robot and a human participant perform the described task. The robot functions according to the same

assumptions as the one described by [87]. The robot acts in two different conditions: It can place the final piece proactively (try to do it itself) or "lazily" (wait until the human places it). During the task, the robot exhibits gaze behavior in accordance with the gaze controller derived from the empirical data collection. There will be two baseline conditions, namely one where the robot does not display any gaze behavior at all, and another one, where the robot acts according to "anti-timings," as in the study of [23].

For the gaze controller, there are numerous possible elaborations. For example, the state space of the temporal roles could be expanded by the belief of who the believed actor of that action is. The state space would then be $\{past, previous, current, next, future\} \times \{robot, human, Xagent\}$. The robot gaze could thus vary when the robot believes that the human is about to perform the next action in contrast to when the robot believes that it is to perform the next action itself.

While the approach in [77], [78], and [23] has worked in conversation settings, it is unclear how gaze processes with dynamic gaze targets are handled by a robot. As human-like object manipulation capabilities are the current goal of service robotics research, human-like gaze behavior in object manipulation tasks is also beneficial, as humans are known to actively seek out information that helps solve the current task. This approach has a counterpart in robotic vision, called *active vision* [33]. Future research can make use of the derived gaze timings to more reliably focus on important aspects of a scene, according to the ongoing task.

3.5 Conclusion

In this chapter, we mainly focused on research in psychology and HRI on joint attention, although there are numerous other related interesting subfields that influence how to think about joint attention in service robotics.

In psychology, attention is studied in numerous different scenarios, such as sustained attention, vigilance, and other low-level models of attention. In developmental psychology, research on the autism spectrum disorder in infants and developmental robotics explore how social collaboration abilities develop and emerge in complex behavior from more simple prerequisites. Studies in neuroscience and psychophysics focus on the neurological processes leading to the attention phenomenon. Differential psychology studies how personality traits lead to different modes of attending to stimuli.

Similarly, for AI/robotics, there are numerous fields that deserve a mention in attention research. Visual attention is an inductive bias, often used in visual pattern recognition and machine learning research. Multi-agent reinforcement learning deals with the emergence of communication protocols between untrained agents and how they attend to each other to solve complex collaborative tasks. In different computational cognitive architectures, joint attention may be a feature that emerges from the dynamic interplay of different architecture components. In machine vision, object detection plays a critical role regarding which objects can be paid attention to. Only if an object is detected, segmented, or classified, it will be able to enter the center of attention. In planning and scheduling, there are numerous different paradigms with many different frameworks, of which a single one was chosen as the focus in this chapter.

To summarize this chapter, first, structural and procedural models of joint attention from the psychological perspective were discussed. The special relation between ToM and joint attention was of particular interest. We then focused on gaze as the main sensory modality. Information gathered through gaze not only provides necessary information to calculate mental representations of one's surroundings, but it is also driven top-down to focus on areas that are crucial to form a coherent explanation. This gaze behavior can be a source of information for observers.

Second, we reviewed how these insights are used to create robotic implementations for different joint attention or joint action scenarios. The scenarios included conversations with locations of interest other than conversation partners or collaborative physical tasks with different manipulable objects.

Third, decision-theoretic and classical planning were reviewed for their use in such collaborative physical tasks. Special attention was paid to plan recognition and the usefulness of a benchmark (building a tower out of blocks) for joint action in HRI.

Finally, we proposed a method for learning a stochastic gaze controller for such tasks from data. The joint action benchmark of jointly building a tower was used as experimental foundation. We presented a method to preserve the time-invariance assumption of the stochastic controller by assigning temporal roles to objects. These roles are assigned dynamically by checking the current world state and the shared plan. This was followed by an outlook on future research needed for the development of a novel gaze mechanism for joint actions in HRI.

The work presented in this chapter only is a building block to a significantly larger research problem, namely how to enable humans and robots to succeed in dynamic collaborative tasks. However, it also demonstrates that attention is a topic that must not only be considered relevant for HRI research, but for the entire robotics field.

Chapter 4

Plan Recognition from Object Detection Traces

The previous chapter described how to produce joint attention gaze behavior in the time resolution of up to a minute when the collaborators have a shared task and goal in an adequately defined setting. However, service robots that co-inhabit spaces with humans will have to act in a more unconstrained environment with regard to their task domains, different goals within the domains, and time horizon. Several challenges in the time scale of up to several minutes arise when a robot observes a human interacting with objects in the environment. When does a task start? What is the current task domain (e.g., fetching objects, setting the dinner table, tidying up the kitchen)? [124], [125] What is the specific goal in the task domain? And how can a robot help the human once it has identified the current goal? It is desirable that robots recognize when a human is performing a task that lies within their area of capability and subsequently estimate the pursued goal within the respective domain.

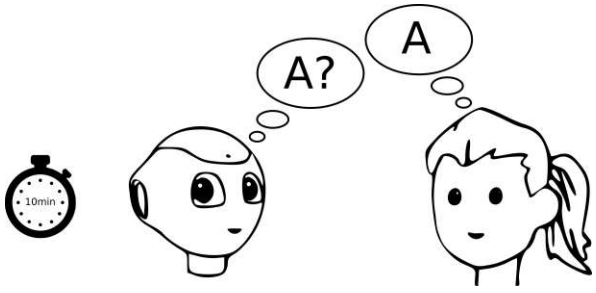

Figure 4.1: In the time resolution of up to ten minutes, robots must recognize human goals and plans during physical object manipulation tasks to formulate helping actions.

In order to fulfill this requirement, a robot must first gather relevant information about its surroundings. This can be done using, e.g., RFID sensors, smart home sensors, or marker-based systems. However, these setups are intrusive and cumbersome to install, preventing their general applicability and acceptance in everyday environments. In this chapter, we rely explicitly on video data, which is much more desirable because sensors are typically integrated on standard robot platforms.

Many household activities have multiple ways to achieve a specific goal. In this thesis, we examine food preparation as a real-life domain relevant to service robotics and include many of the challenges mentioned above. For example, there are different ways of preparing a meal in a fully equipped kitchen. Different utensils fulfill the same function, ingredients vary, and the execution steps are only partially ordered. Thus, human plans to solve long-horizon high-level tasks show much variance. However, an observing robot should still be able to identify the task domain and goal and choose a practical action depending on the chosen plan (Figure 4.1). E.g., the robot may step back when a human is reading a book in the living room but must step in when the human is doing a chore. Robots in long-term deployment will be confronted with human behavior from different domains; not all are relevant to a robot.

Once a goal is identified, the robot must find a way to help the human in the activity, i.e., determine which actions in the planned action sequence are eligible for the robot to perform. Notably, the kind of robot intervention in complex human tasks depends on the desired mode of interaction during collaborative tasks. If the robot reacts to human commands, there are different command-based interaction modalities [126], such as speech-based, gesture-based, and brain-computer interface-based command modalities. More broadly, there are different interaction styles for collaborative tasks. [7] identify different variations of autonomous (autonomous, proactive, reactive), human-led (human help request, human command), and robot-led (robot commands the human what to do, the robot provides information about what the human could do) interaction styles.

For most of these different interaction styles, some underlying capabilities must be developed: detecting the world state, predicting the actions of the collaborator, and formulating plans for their path of action.

Service robots in long-term deployment already find high-level plans to solve complex tasks that take several minutes to execute. High-level automated planners are used to arrive at single-agent plans in their designed task domain. In previous work, it has been shown that this single-agent task knowledge can also be used for plan recognition [112] while observing other agents under the assumption that the observed actor indeed acts in the estimated task domain.

A robot should not only recognize a goal, but also be able to reason about the observed human action and formulate its own plans (e.g., to help accomplishing an ongoing task) from an image sequence. However, automated planners rely on an abstract representation of a problem, e.g., using the PDDL planning language formalism. Robots, on the other hand, only receive time series of sensor readings. A domain-specific state filter must process the sensor readings to provide a formulation of the world state to the automated planner. For example, a knife touching a carrot could mean that the carrot is cut. Therefore, the fact `isCut(carrotA)` can be added to the current state in the automated planner.

Symbolic planning paradigms like classical planning, answer set programming or constraint satisfaction programming possess the necessary abstraction level to solve different types of problems. Among these, classical planning uses the most natural formulation for agent-centric problems.

Action recognition [127] can be used to determine specific actions. In robot-centered task domains, however, manipulating objects is often the focus. In such object-centered

physical tasks, it has been found that the sequence of physical contact between objects holds information to identify actions. In [128], [129], the authors use this idea to arrive at an unsupervised categorization of different actions. However, it is not enough for robotic tasks to identify an action in isolation from the rest of the ongoing activity. It must be clear which objects are involved and how the detected action is used for plan recognition. Automated planning formalisms allow defining actions that take arguments. The arguments can be objects of the planning problem instance. We use the idea of [129] and apply it as a state filter to an automated planning problem, thus allowing us to perform plan recognition merely from 2D object detections. We show that without explicit action recognition, only with 2D object detection available can we prune the many possible plan hypotheses and estimate the current goal. An advantage of this method is the usage of planning capabilities that a robot needs to solve a task independently. Thus, improvements in the robot's planning capabilities will also positively affect its plan recognition capability. We provide service robots a way of estimating goals from sensor data in the time scale of several minutes of human activity that fall into relevant robotic task domains and allow them to help the human.

The symbolic tokens in classical planning problems must be instantiated from subsymbolic data like pixels in a video. Lifting such data to a higher level poses several problems. To avoid operating with meaningless symbol tokens, [105] proposes a bottom-up representation of two parts: an iconic representation, which can detect object instances in sensory input (e.g., the many shapes an object can cast on the retina of an observer) and a categorical representation, which holds information about class invariant features. In this chapter, we substitute object detection algorithms and knowledge bases for these two components.

We propose a method to infer goals and plans from an image sequence given object detections, a PDDL domain and instance, an object ontology, and a set of goals. A knowledge base is used to complete affordance and object property knowledge in a given PDDL instance template. An object detection algorithm provides bounding boxes of objects in the video sequence. A sequence of sets of possible planning actions is created from the completed domain model and sequence of object interactions. From this sequence and together with the planning instance, a Monte-Carlo-Tree-Search (MCTS) procedure creates a directed acyclic graph (DAG), where nodes represent states and edges represent actions. This search graph can represent all possible plans that can be generated from the sequence of possible planning actions. Using the properties of MCTS and using a plan recognition algorithm as a rollout policy, parts of the search graph that lead to a goal are predominantly expanded.

The choice for using a MCTS search graph results in a drastically reduced length of the observation trace used in each call to the rollout policy. Using the given plan recognition algorithm directly on the whole sequence of possible actions usually does not terminate or find a solution. This is due to the fact that there are usually much more object interactions in a video sequence than necessary planning actions to fulfill the corresponding planning problem. Also, depending on the problem formulation, an object interaction can often instantiate two planning actions with complementary effects, such as "pick up" and "put down".

Current evaluation uses the annotated dataset in [130], which is a portion of the

larger, but not as thoroughly annotated, MPII 2 Cooking dataset [131]. We use classical planning domain formulations in two different degrees of complexity. Due to partial observability and multiple cooperative agents (e.g., two hands that need to work together to achieve some effects), this is a challenging planning domain.

In summary, this chapter makes the following contributions:

1. a novel problem formulation of goal estimation procedure from video without explicit activity recognition,
2. plan representation through DAG,
3. analysis of MCTS evaluation schemas for goal selection.

The chapter is organized as follows: Related work is presented in the following section (Chapter 4.1). We next provide a detailed description of each step in our proposed method (Chapter 4.2). Next, an evaluation and different proposed metrics are discussed (Chapter 4.3). This is followed by an outlook (Chapter 4.4) and a conclusion (Chapter 4.5). The content of this chapter is based on previously published work in [132].

4.1 Related Work

4.1.1 Plan Recognition

[113] describe the different layers of the recognition problem and the different approaches within each area. They summarize that activity recognition and plan recognition form a pipeline from low-level sensor to high-level complex goals and intents. [133] argue that activity and plan recognition are conceptually inter-related by formalizing the general recognition problem.

Most relevant for our work, [26], [112], and [114] propose plan recognition procedures with increasingly relaxed preconditions. In these approaches, a goal recognition problem is compiled into a classical planning problem that leverages different kinds of classical planning solvers. In another approach, [134] develop a goal recognition procedure using partially observable Markov decision process (POMDP) planners for the stochastic case, where the agent and observer have access to the same POMDP model.

Further results, elaborating the plan recognition as planning approach, are given by [135], and [136]. The worst-case distinctiveness of a plan recognition problem is a measure that describes the maximal length of a plan an agent pursues before being able to deduce the actual goal. This measure can be used to create domain models where agents reveal their true goal as early as possible. [137] define the necessity of a proposition for a set of goals in a given plan recognition instance, which describes the percentage of goals that require the said proposition to be true. This measure can be used to detect helpful actions in a multi-agent scenario.

4.1.2 Activity Recognition

Many methods to infer current human activity from time series data like videos or human pose estimations have been proposed [138]. [139] comprises methods and methods to infer plans from a series of atomic actions. Plan recognition approaches are said to mostly operate under the keyhole paradigm, where the unobserved agent is unaware of its observer, which is the case in this chapter, too.

[128], [129] demonstrate that object interaction, in the case of contact between two objects, can describe activities. In contrast, we study whether object interactions hold enough information to directly solve symbolic plan recognition problems.

The prior work proposes Semantic Event Chains (SEC) for manipulation actions, which capture changes in the spatial relations between a manipulator (e.g., human hand or robotic gripper) and manipulated objects. A strength of this approach is that the object properties do not play a role in the identification of actions. This approach enables unsupervised learning, where clustering of different types of SECs enables different manipulation sequences to be learned, e.g., hiding an object will result in a different SEC than cutting an object with a knife.

This is an inspiration for our approach. We assume that in order to solve a symbolic planning problem with planning operators of arity 2, a sequence of planning actions instantiated with a task-specific combination of objects present in the planning instance must be generated from a video.

4.1.3 Knowledge Representation

Regarding a more general issue, [105] formulates the symbol grounding problem, which is especially relevant for robotic contexts, as shown in the review by [106]. [140] propose a learning-from-demonstration algorithm that can learn symbolic representations of sensory input patterns through the intermediate clustering in conceptual spaces.

There have been several approaches to organizing relational knowledge [141], and especially robot-centered knowledge [125] [142] in knowledge bases. These knowledge processing systems support storing and reasoning on semantically annotated subsymbolic data and use common-sense ontologies for the given everyday use-case.

4.1.4 Plan Estimation from Video

[143] devise an approach that estimates goals from video frames and plan libraries. They use a CNN for explicit activity recognition, creating a sequence of activities for each input video. This sequence is then matched to sequences of cooking actions. Each recipe is represented as a tree of cooking actions. This approach is able to perform goal recognition from video, however, it does not allow individual objects in the plan to be considered.

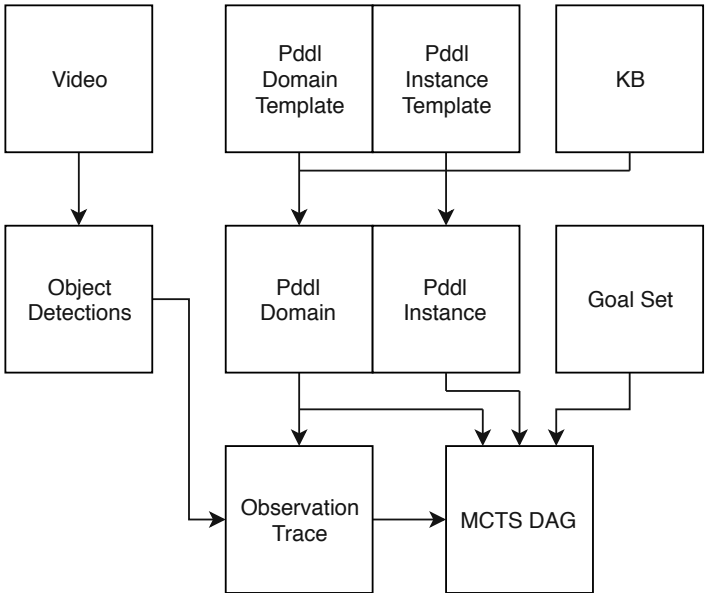

Figure 4.2: System overview. From a video sequence, the spatial object annotation is created. A PDDL template is completed by a knowledge base. The observation sequence results from the object interaction list and the PDDL instance. An MCTS procedure estimates a posterior goal probability for the goal set.

4.2 Method

This method is developed for sufficiently complex domains where a single object interaction cannot automatically achieve a goal. Appropriate problems contain goals that are achievable through various partially ordered plans. This section details the building blocks of the approach (Fig. 4.2).[1]

4.2.1 Object detection

From a video, a spatial annotation of relevant objects is created, provided by methods like [144] or [145]. With such an annotation, the object interaction sequence $\mathcal{O}'$ is composed. Under the same assumption as in [128], the start and end point in time of two objects having physical contact holds information about the whole sequence of actions within a given sequence. $\mathcal{O}'$ is a list with strict ordering by frame number. For each frame, a pair of objects are added to a set when the interaction between them begins or ends according to their spatial annotation. We could also define a start and a stop set per frame to use all the provided information. If a planning action is matched to an object interaction, it is then assigned either to the start or stop set, which must be chosen by the domain designer. This is left for future work. For online use of our approach, fast object detection systems need to be pre-trained with all objects relevant to the domain.

[1] Code available at `https://github.com/michaelkoller/pic-to-plan-v2-git`

4.2.2 Classical Planning and Plan Recognition

In [112], a STRIPS planning domain is a tuple $P[\cdot] = \langle F, I, A \rangle$, where F is a set of fluents, $I \subseteq F$ is the initial goal state, and $G \subseteq F$ is the goal state, which completes $P[G]$ to a planning problem. A is a set of actions, where $a \in A$ are actions with $Pre(a)$, $Add(a)$ and $Del(a)$ as its preconditions, add and delete lists.

A plan recognition problem is a tuple $T = \langle P[\cdot], \mathcal{G}, O \rangle$, where P is a planning domain and $\mathcal{G}$ is a set of goals $G \subseteq F$. $O = o_1, \ldots, o_m$ is the observed action sequence, where $o_i \in A$, $i \in [1, \ldots, m]$. More specifically, a probabilistic plan recognition problem is a tuple $T = \langle P[\cdot], \mathcal{G}, O, Prob \rangle$, where $Prob$ is a probability distribution over $\mathcal{G}$. For this work, we adopt the probabilistic formulation and change the definition of the observation trace: $\mathcal{O} = \{o_{11}, \ldots o_{1p}\}, \ldots, \{o_{m1}, \ldots, o_{mq}\}$ is a strictly ordered sequence of sets of actions. Our plan recognition problem is thus a tuple $T' = \langle P[\cdot], \mathcal{G}, \mathcal{O}, Prob \rangle$. The result of our plan recognition problem is a posterior probability distribution over the set of goals.

4.2.3 Knowledge Base

In order to keep domain models and instances consistent and more manageable, we define a knowledge base per planning domain and instance, which completes the given domain and instance template. The ontology $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, where the TBox $\mathcal{T}$ is a finite set of general concept inclusions ($C_1 \sqsubseteq C_2$) and the ABox $\mathcal{A}$ is a finite set of concept and role membership assertions ($C(a), R(a,b)$). Using consistency checks and automatic inference reduces the effort for the designer.

Individuals defined in the knowledge base are assumed to be present in the scene. This can include past observed objects, objects that are usually present given a certain context (e.g., in a functional kitchen, there will be a sink), and hypothetical objects. Similar to [146], hypothetical objects are objects in the database that incur a higher cost than regular objects when being used in planning operators to achieve a goal. Therefore the planner will avoid using hypothetical objects (e.g., prefer retrieving an observed tool and using it instead of a hypothetical one). However, if the observed object interactions that constitute a planning problem strongly suggest the use of an unobserved hypothetical object, it might be that the robot lacks relevant knowledge of the current observed scene. Currently, it is up to the designer to define which typical objects are hypothetical.

4.2.4 PDDL Domain and Instance Completion

In order to complete the PDDL instance, concepts declared in $\mathcal{T}$ are inserted into the domain definition, whereas all concept and role memberships of individuals (i.e., fluents) occurring in $\mathcal{A}$ are inserted into the initial state description. All individuals are inserted into the objects definition of the planning instance. This way, only predicates that are necessary for the intended formulation of the planning operators need to be explicitly defined in the planning domain model.

4.2.4.1 Compilation of PDDL Domain

As a result of the binary interactions of the interaction sequence $\mathcal{O}'$, all planning operators must also be defined with arity of 2. This is often a natural fit for manipulation-heavy domains, but sometimes refactoring is necessary. In our formulations, we are always able to compile operators of arity greater than two into binary operators with the method described below.

To show this, we use unary and binary predicates since predicates of higher arity can be refactored to semantically equivalent groups of predicates with lower arity. For a domain $\mathcal{D} = \{d_1, \ldots, d_m\}$ and an n-ary predicate P, introduce m^n predicates of the form $P_{x_1,\ldots,x_{n-1}/x_n}$, with $x_i \in \mathcal{D}, i \in [1, \ldots, n]$. The intended semantics of this is shown with the example $\mathcal{D} = \{u,v,w\}$ and a model I, where $I \models P(u,v,w) \Leftrightarrow P_{uv/w}(u,v) \wedge P_{uw/v}(u,w) \wedge P_{vw/u}(v,w)$.

Planning operator definitions of higher arity can be reformulated to lower arity under the assumption that certain predicates represent a resource, e.g., a hand can only hold one object at a time.

```
(:action A_3
  :param (?x ?y ?z)
  :precondition (and a(x) b(y) c(z)
    d(x,y) e(y,z) f(x,z))
  :effect (and a'(x) b'(y) c'(z)
    d'(x,y) e'(y,z) f'(x,z))
)

(:action A_2
  :param (?x ?y)
  :precondition (and a(x) b(y) d(x,y)
    (exists (?z) (and c(z) e(y,z) f(x,z))))
  :effect (and a'(x) b'(y) d'(x,y)
    (forall (?z) (when (and c(z) e(y,z)
    f(x,z)) (and c'(z) e'(y,z) f'(x,z)))))
)
```

It is now also necessary to define an operator that can affect the fluents containing the factored variable, in this case z and $c(z)$, $e(y,z)$, $f(x,z)$, which are also of lower arity. The designer has to impose an order on the refactoring sequence.

4.2.4.2 Observation Trace

Next, we want to instantiate actions that can result from an object interaction, comparing static preconditions of planning operators and static object properties defined in the knowledge base. Therefore we define a mapping $f : \mathcal{O}' \rightarrow \mathcal{O}$ from the interaction sequence $\mathcal{O}'$ to the sequence of possible actions $\mathcal{O}$ using the completed domain model. For each interaction between $\{o_1, o_2\}$ both permutations $(o_1, o_2), (o_2, o_1)$ are considered. For an object interaction tuple (p,q), all object concept memberships $C_i, i \in \{p, q\}$

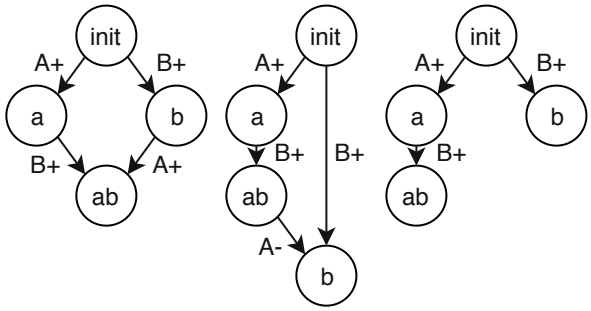

Figure 4.3: DAGs for $\mathcal{O}_1 = \{A+, B+\}, \{A+, B+\}$ (left), $\mathcal{O}_2 = \{A+, A-\}, \{B+, B-\}, \{A-\}, \{B-\}$ (middle), $\mathcal{O}_3 = \{A+, A-\}, \{A+, A-\}, \{B+, B-\}$ (right).

are queried from the knowledge base, and for each planning operator, preconditions $P_i, i \in \{p, q\}$ that also occur in the knowledge base are considered. The action $a(p,q)$ is instantiated if $P_p \subseteq C_p$ and $P_q \subseteq C_q$. This represents a necessary condition for actions to be able to arise from an object interaction. This way, many object interactions are filtered. However, for a single object interaction, many actions can be instantiated, e.g., if there are multiple matching actions (`(Stack A,B)`, `(Unstack A,B)`), or both arguments are from the same object class (`(Stack A,B)`, `(Stack B,A)`).

4.2.5 Monte Carlo Tree Search Directed Acyclic Graph

The generated list of possible actions $\mathcal{O} = O_1, \ldots, O_m$, with $O_i = \{o_1, \ldots, o_n\}$ is the observation trace of our plan recognition problem. In our experiments with the given domain model, there are usually hundreds of sets of actions and $|O_i| > 1, i \in [1, \ldots, m]$.

4.2.5.1 Search Graph

From this list, a directed acyclic search graph can be built, where nodes represent states in the planning problem, the root node represents the initial state from the PDDL instance, and directed edges represent actions that transform the origin state to the destination state. If a node has multiple incoming edges, there exists a lowest i among the sets O_i to which the incoming edges belong. All actions of all sets $O_j, j > i$ must be considered as possible children of the current state. If an edge would create a cycle in the graph it is not inserted. Cycles would be created if an action transforms the current state in a previous state such that a directed path from the resulting state to the current state already exists in the DAG. This excludes only suboptimal plans from consideration. For example, consider the toy planning domain with the fluents a,b and the actions `A+`,`B+`, `A-`,`B-`, which add or delete the respective fluent to the successor state. Fig. 4.3 shows examples of DAGs created from different $\mathcal{O}$.

Algorithm 1 creates the DAG (cf. [147]). With some abuse of notation, we equate states to nodes. VAL [148] is used to evaluate if an action is applicable in a given state, as it is not guaranteed that all preconditions are given.

The resulting DAG represents all possible sequences of actions given an $\mathcal{O}$, and its

Algorithm 1 DAG Construction

```
1:  procedure ConstructDAG(O, n_init, δ)
2:      N ← {n_init}                                   ▷ Node set
3:      E ← {}                                         ▷ Edge set
4:      for i = {1, ..., |O|} do
5:          N' = {}
6:          for n ∈ N do
7:              for a ∈ O_i do
8:                  n' ← δ(n, a)                       ▷ Apply a to n
9:                  if n ≠ n' then
10:                     if n' ∉ N then
11:                         N' ← N' ∪ n', E ← E ∪ (n, n')
12:                     else if ∄(n' ⇝ n) then         ▷ Avoid
13:                         E ← E ∪ (n, n')            ▷ cycles
14:         N ← N ∪ N'
15:     return N, E
```

depth is bound by $|\mathcal{O}|$, where only few paths down the DAG represent goal-directed behavior, if at all. Monte Carlo Tree Search is well suited for informed asymmetric tree expansion. Additionally, using MCTS is an anytime algorithm, and observations can be added to the observation trace incrementally.

We use adaptations presented in [147]: In a DAG, $c(x)$ denotes the set of edges going out of x, if x is a node. If x is an edge, then $c(x)$ denotes the set of outgoing edges of the destination of x. For an edge x and its origin y, $b(x) = c(y)$, the set of siblings of edge x, including x. $\mu(x)$ and $n(x)$ denote the mean reward and the number of visits to the respective node or edge. $p(x) = \Sigma_{e \in b(x)} n(e)$ denotes the total number of payouts to the sibling set of x. $\mu'(x)$ and $n'(x)$ denote the reward that is attached to an object x, which is attached to it between its first appearance and the first appearance of a child of x. [147]

$$\mu_0(e) = \mu(e) \tag{4.1}$$

$$\mu_d(e) = \frac{\mu'(e) \times n'(e) + \Sigma_{f \in c(e)} \mu_{d-1}(f) \times n(f)}{n'(e) + \Sigma_{f \in c(e)} n(f)} \tag{4.2}$$

$$n_0(e) = n(e) \tag{4.3}$$

$$n_d(e) = n'(e) + \Sigma_{f \in c(e)} n_{d-}(f) \tag{4.4}$$

$$p_d(e) = \Sigma_{f \in b(e)} n_d(f) \tag{4.5}$$

This results in the UCT score formula $u_{c,d_1,d_2,d_3}(e) = \mu_{d_1}(e) + c \times \sqrt{\frac{\log p_{d_2}(e)}{n_{d_3}(e)}}$ with $(d_1, d_2, d_3) \in \mathbb{N}^3$ to determine the descent path in the tree policy. In our experiments, we used $d_1 = \infty, d_2 = 0, d_3 = 0$.

4.2.5.2 Rollout Policy

The probabilistic plan recognition procedure presented in [112] is then used as the rollout policy using a satisficing configuration of the FastDownward planner. Using satisficing instead of optimal planners results in faster solution generation and is alleviated to a certain degree by our parallelization method.

In each rollout policy call, a probabilistic plan recognition problem $T = \langle P[\cdot], \mathcal{G}, O, Prob \rangle$ is solved. Typically the length of the used observation trace O, which represents the descent path from root to expansion node, is much shorter than the original observation sequence $\mathcal{O}$. The result of the rollout policy is a probability distribution over the set of goals $\mathcal{G}$. This is stored within the respective edge, and the highest resulting probability value $n' = max(Prob)$ is backpropagated along the descent path. This leads to a tree expansion that favors paths leading to at least one of the given goals, but it does not detect if a path favors multiple goals. This behavior could be accomplished if the normalized sum of probabilities in $Prob$ were backpropagated.

In the case where an edge between two existing nodes is inserted, no rollout is performed. It is, therefore, necessary to copy the n' value and the probability distribution of the existing in-going edges of the destination node into the newly added edge.

4.2.5.3 Parallelization

Modern processors enable parallel rollout policy calls to be made. We modify the UCT-MCTS algorithm given a number of processor cores. Once the tree policy has chosen an edge to expand, as many unexpanded siblings as possible with respect to number of cores are expanded simultaneously. Remaining cores are used to recompute the rollout policy on a random selection of edges on the current descent path. If a plan with a lower cost is found, the goal probability distribution and n' value are updated. This alleviates the problem of exceptionally bad planning costs in previous steps. Most of the runtime is spent performing the rollout policy, which prevents the system from real-time application. Efficient domain design and more rare use of the rollout policy during graph expansion might alleviate this problem.

4.3 Evaluation

4.3.1 MPII Cooking 2

[130] present a detailed bounding box annotation for a subset of the MPII Cooking 2 dataset [131]. We use the bounding box object annotation to create the initial touch events sequence. This sequence, together with the ontology and the domain, comprises the possible actions for an instance. The number of redundant actions is very high and represent a realistic use case. Although, due to the small number of annotated videos and only two different performed recipes, a thorough evaluation is not possible. The complete MPII Cooking 2 dataset is still very suited for our experiments since it contains many different sequences of preparing dishes (goals), variation for reaching

a goal, and significant overlaps in plans for different dishes. Ideally, the whole MPII Cooking 2 dataset with object annotation from [130] would be used for evaluation.

An evaluation is performed on the dataset of [130], which consists of 9 annotated video sequences. Each depicts a person preparing utensils and ingredients to either cut a loaf of bread (4) or a cucumber (5) on a kitchen counter. For these instances, also simple classifiers would work.

4.3.1.1 Domain definitions with increasing complexity

The kitchen setting is a suitable domain for our approach, where many variations of object manipulations can result in the preparation of many different dishes. For evaluation purposes, we use two planning domain formulations with increasing complexity that are inherently multi-agent because there are two hands in the scene that can perform different actions. In the more detailed formulation, some actions also require cooperation between the two agents, e.g., in order to cut an ingredient, it must be grasped with one hand and cut with a knife held in the other hand.

4.3.2 Baseline Comparison

The observation trace $\mathcal{O}$ is a list of sets of actions, strictly ordered by frame number. One can transform this observation trace into a regular observation trace like in [112] by choosing any ordering within each frame set and then flattening the list of sets. As a baseline comparison, the regular probabilistic plan recognition algorithm can then be applied to solve the problem. In our tests, the observation traces contained hundreds of actions, and the planner was not able to find plans for the transformed planning problems. Therefore all goals were reported to be equally likely. In contrast to this, in our approach the default policy calls usually contain only a small number of observations and terminate within the specified search time limit. [114] propose a compilation that allows observations to be discarded at a parameterized cost, which could alleviate the problem of domains with complementary actions like `Stack` and `Unstack`, but it is unclear whether the general problem of very long observation traces is handled more successfully than in [112].

4.3.3 Evaluation Metrics

Typically, MCTS is used to evaluate which action to take from the root node in order to maximize some reward between 0 and 1. We operate with a probability distribution as reward instead of a single numeric value. Therefore we define new metrics for this approach. Given a graph $G = (N,E)$ of nodes N and directed edges E, different metrics to determine the most likely goal can be defined. $Prob_e$ represents the probability distribution over the goal set $\mathcal{G}$ stored in edge e. Note, that each node shares the same $Prob$ with the in-going edge of the highest probability.

Metrics can be defined with respect to the considered node and edge set:

- descent path from root node choosing the highest reward, $\mu_d(e)$ at each edge (best goal can change along the path),
- descent paths from root node choosing the highest reward, $\mu_d(e)$ per goal $g \in \mathcal{G}$,
- whole edge set E, or whole node set N,
- in-going edges of leaf nodes in N or leaf nodes in N.

Descent paths can terminate as soon as an actual goal state or a leaf is reached. Next, in the specified edge set $E' \subseteq E$, the most likely goal can be defined according to the

- highest value in a single edge, $\arg\max g \in \mathcal{G} max_{e \in E'} Prob_e(g)$,
- largest sum over the edge set, $\arg\max g \in \mathcal{G} \Sigma_{e \in E'} Prob_e(g)$.

Table 4.1 presents results for a selection of the above-defined metrics in the more complex domain formulation used. For the different metrics, there is a tendency that sums of probabilities over the whole or the leaf set of nodes achieve the highest accuracy. Using edges was not advantageous, i.e., counting nodes with multiple incoming edges did not show any improvements. Among the sum metrics, there also is the tendency that the metrics show a greater relative difference the more of the observation sequence is used as input (first 10%, first 50%, 100%, Fig. 4.4). The maximum value metric seems promising and should first be evaluated in more complex datasets.

Similar results arise in the less complex domain definition, except that no classifier reaches accuracy of 1. This does not lead to the conclusion that a more complex domain model helps in the goal estimation, but it will be a line of inquiry for the full dataset evaluation. As above, the sum metrics generally achieve better accuracy, with the descent path metric being highest (0.9).

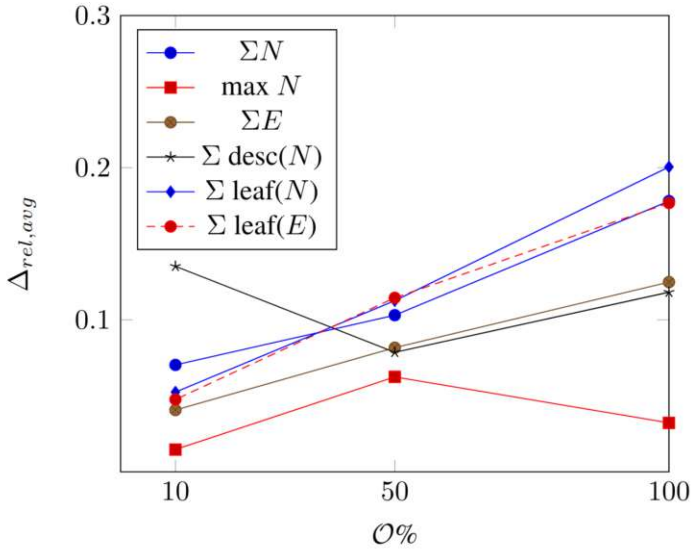

Figure 4.4: The plot depicts how the average relative difference between true and false label score changes with respect to the length of the observation trace, for several metrics in the complex domain. The bigger the difference, the better a metric can distinguish between classes.

$\mathcal{O}$%	Metric	avg. score true / false label	acc.
100 %	ΣN	$20.7 \pm 10.5/14.4 \pm 7.9$	1.0
	Σ leaf (N)	$11.2 \pm 5.3/7.4 \pm 4.4$	1.0
	max N	$0.88 \pm 0.2/0.83 \pm 0.2$	0.55
	Σ desc(N)	$1.6 \pm 9.5/1.2 \pm 0.6$	0.66
50 %	ΣN	$20.8 \pm 6.9/16.9 \pm 8.2$	1.0
	Σ leaf (N)	$11.4 \pm 6.9/9.1 \pm 7.2$	1.0
	max N	$0.94 \pm 0.16/0.83 \pm 0.25$	0.61
	Σ desc(N)	$1.7 \pm 0.6/1.4 \pm 0.8$	0.83
10 %	ΣN	$31.4 \pm 14.8/27.3 \pm 16.5$	1.0
	Σ leaf (N)	$16.6 \pm 10.7/14.9 \pm 11.2$	1.0
	max N	$0.88 \pm 0.22 \pm 0.86 \pm 0.27$	0.55
	Σ desc(N)	$2.1 \pm 0.8/1.6 \pm 0.8$	0.66

Table 4.1: Results on MPII Cooking 2 dataset fragment ($n = 9$) with 2 dishes (sliced cucumber/bread). Columns are divided into percentage of observation trace $\mathcal{O}$ used ($\mathcal{O}$ %), the used metric, the average score for the correct and incorrect predicted label), and accuracy (acc) of the used metric. The used metrics are $\arg\max_{g \in \mathcal{G}} \Sigma_{n \in N} Prob_n(g)$ (ΣN); $\arg\max_{g \in \mathcal{G}} \Sigma_{n \in N'} Prob_n(g)$, $N' \ldots$ leaf nodes (Σ leaf (N)); $\arg\max_{g \in \mathcal{G}} \max n \in N Prob_n(g)$ (max N); and $\arg\max_{g \in \mathcal{G}} \Sigma_{n \in N'} Prob_n(g)$, $N' \ldots$ best descent path (Σ desc (N)).

4.4 Outlook

Twin domain definitions If a robot is to reason about the goals of an observed human, it is desirable that it has the means to formulate assistive plans toward the estimated goal. Therefore we assume that a robot and a human will accomplish tasks in a similar fashion in object manipulation domains (i.e., gripper and hand). However, a human will typically have more flexibility in reaching a goal. This can be expressed by less strict preconditions of the involved planning operators. The PDDL domain model for a certain task for a service robot can be used to create a more suitable PDDL domain model for observing humans. Heuristically, a human observation domain model can be generated by a human designer by relaxing the preconditions of planning operators and removing the then unused predicates.

PDDL extensions The PDDL formulation of an instance matched better to the given scenario by the explicit formulation of partial observability. [149] propose a PDDL compilation procedure using the knowledge operator K that transforms a fact p into two facts of the form Kp and$\neg Kp$. Thus, the state of knowledge about whether p is believed to be true or false can be reasoned about.

Using planners that can handle domains with durative actions would allow to exploit time interval information inherent in the observation sequence. Additionally, the definition of durative actions could be leveraged to filter out noisy object interactions.

Missing Observations The presented formalism cannot cope with missing observations. This is due to the strict correspondence between nodes and states, as well as between edges and actions. Including missing observations would result in a node representing all planning states containing the fluents of a node. The MCTS can be adjusted to incorporate missing observations by not using VAL to perform validity checks on single actions between two consecutive states but by finding valid plans between two observed states. A plan with only one action represents the original case of only one action between two observed states. In the case of a single missed observation, the planner will find a plan with two actions, where the first action leads to the missed state and the second actions leads from the missed observation to the original second state. By adjusting the parameter of how long the plans between two observed states is allowed to be, a number of consecutive missed actions can be accounted for. The trade-off is a larger search space and longer computation times. The allowed number of missing observations between two observed states can also be probabilistically sampled from a distribution. Reasonable sampling distributions should resemble the error distribution of missed observations.

4.5 Conclusion

The evaluation provides results for the problem of inferring goals from video using spatial object annotation traces. This is done by matching the properties of detected objects to the preconditions of planning operators, using a knowledge base. The resulting observation trace contains significant noise, which is handled by compactly representing all possible action sequences in a DAG. A variant of MCTS leveraging a previous

plan recognition algorithm as rollout policy is applied for informed expansion of the search graph. Thus, we have developed an approach that does not rely on an explicit activity recognition step to instantiate a plan recognition instance. Instead, we directly transform object interactions into classical planning actions. An evaluation provides insights of which defined metrics work best for the adapted MCTS.

Chapter 5

Generating Well-Annotated Object Interaction Samples

In this chapter, we stay in the same time resolution as the previous chapter and are still concerned with action sequences of plans that can be completed in the time span of a few minutes. However, we turn our focus from algorithms to datasets.

Research in robotics depends on suitable datasets for the specific research topic to speed up training or testing time and standardize benchmarks. For example, image datasets of everyday objects with ground truth annotation for lower-level robotic actions such as grasping are widely used (e.g., YCB-Video [150]). However, joint attention is a problem outside of a static context. There also exist time series datasets, for example, video datasets that depict humans performing everyday actions (e.g., MPII Cooking 2 Dataset [131]). Extensive ground truth annotation in such datasets is prohibitively expensive, which limits their usefulness.

Service robots in human homes will be expected to perform menial tasks in households and to cooperate with their human partners. Thus, HRI researchers have studied kitchen settings since they combine many challenges like collaborative object manipulation [151], [152], use of semantic knowledge in interaction tasks [153], and plan recognition [154], among others. Kitchen environments are also challenging environments in HRI studies concerning proxemics and social navigation [155].

Datasets about joint attention in everyday activities have several requirements. In each sample, many interactable objects should be available to the actor. Many object classes and multiple instances of an object class should exist in the scene. The dataset should include different goals for an actor with different ways of achieving one goal. The dataset should include different typical robot sensor data types, such as RGB and depth images, as well as segmentation masks. In the ideal case, a dataset should include a complete ground truth annotation for all sensor data. For video samples with several minutes each, the ground truth of the world state should also be annotated at an appropriate abstraction level. This is achieved by including high-level object predicates specific to everyday household tasks. These predicates describe the relationship between objects, such as which objects rest on top of another, are stored inside of another, and so on. There might also be the need for additional logical predicates for a specific domain. For example, in a kitchen, the stove can be turned on or off, the fridge or drawers can be open or closed, and so on.

Figure 5.1: Dataset example for preparing salad. Left: RGB view; middle: depth view; right: object mask view.

One approach is to record real-life samples and then use human labor for ground truth annotation [130]. Another practice is to apply current state-of-the-art solutions to arrive at a ground truth annotation. However, annotation errors will then be interpreted as ground truth in later use. Some annotation types might be automatically generated by installing additional sensors in the recorded scene, e.g., distance sensors or RFID tags for detecting open drawers.

In recent years, synthetic datasets have been used to alleviate these problems [156], [157]. Their main benefits are the correctness of the ground truth data and the ease of data generation and annotation. It has also been shown that algorithms trained on simulated data perform well in real-world settings. Overcoming the so-called "sim2real gap" [158] has recently gained much research interest. For our use case, however, we are interested in the data quality and sample complexity to create a challenging test environment for robotic joint action tasks.

We present the *Virtual Annotated Cooking Environment* (VACE) (Figure 5.1), a new video dataset and Unity-based virtual simulator in a rich kitchen environment, modeled after MPII Cooking 2 [131], which is a well-known existing dataset for activity recognition of fine-grained and composite actions showing different persons performing various cooking tasks. This dataset aims to facilitate research on activity and plan recognition, learning from demonstration, and semantic segmentation research. One of the main advantages of virtual environments is the quick and reliable annotation of the generated samples because of the accessibility of virtual ground truth information, thus generating ground truth annotation for free. Users easily record data samples while using standard virtual reality equipment. Samples are automatically recorded and rendered after the recording phase. In our work, all labeling is done by the simulator. The user only has to provide a high-level description of the sample in the readme file. Especially with robotic applications in mind, virtually generated data and virtual simulators can save time while still generating useful insights [159]. In domains shared between humans and robots, activities depicted through human demonstrations have been shown to shorten training time for robotic applications [160] in the learning from demonstration paradigm.

In other fields of interest like semantic modeling or plan and activity recognition, research depends on accurate ground truth for different parts of the computational pipeline from pixel-wise object segmentation to semantic predicates of objects (*object held*, *contained*, etc.), and the ground truth plan or temporally segmented activity label. Especially approaches like [143] can leverage well-annotated datasets that provide all

information along their pipeline from video to logical predicates, since they propose domain-specific plan recognition in a kitchen setting from video sequences. There already exist multiple datasets for activity recognition in everyday human activities, such as cooking or performing other household chores [161], [162]. Also there are multiple virtual environments for plan and activity recognition [156], as well as robotic manipulation tasks [163].

However, we see a need for a dataset that fills a current gap. Ideally, this dataset will include samples with many different goals but overlapping plans to create more challenging instances. Agents in VirtualHome [164] can choose among many goals, but since they are spread over all rooms of a typical apartment, the plans often diverge in the first steps, which makes them easy to classify correctly. VRKitchen [156] on the other hand, models five dishes very intricately with object state changes, but a larger goal set is desirable. The kitchen room in iGibson2.0 [165] is sparsely furnished and does not allow many different recipes.

Two recent surveys of RGB-D video activity datasets summarize the different kinds of scenarios that are currently available for activity and plan recognition research [166], [167]. Since our focus is on understanding long-term semantic plans from video, we are interested in datasets that show behavior composed of multiple consequent actions. Therefore datasets containing samples with only one action such as *sit down*, *exit the room*, or fitness exercises are not suitable. Furthermore, there must be a wide range of possible goals at the beginning of a video sample. For instance, the CAD120 dataset [168] contains sample classes like *making cereal*, and a cereal box in the first frame already gives away the true label.

A very challenging dataset that avoids these problems is the MPII Cooking 2 dataset [131]. It contains samples of many different cooking recipes, in many different variations with respect to plan ordering, used utensils, and ingredients. As an example, when a knife is taken out of a drawer, only a fraction of the recipes is excluded and the true goal could have been accomplished with a different tool as well. The samples in the MPII Cooking 2 dataset are suitably complex, overlapping, and varied. However, the dataset is unfortunately not comprehensively annotated, which is necessary for proper quantitative evaluation of methods. Optimally, each sample in a dataset consists of realistic and complete sensor data, complete semantic and spatial segmentation as well as temporal task-specific semantic annotation. [27] annotated a part of the MPII Cooking 2 dataset (10 samples, 2 different goals) with a suitable level of detail but reports that such annotation constitutes a significant effort. With our virtual reality (VR) simulator, we can recreate these samples and improve their annotation for a virtual environment.

The rest of the chapter describes the dataset (Chapter 5.1), and the simulation environment (Chapter 5.2), and ends with a conclusion (Chapter 5.3). This chapter's content is based on previously published work in [169].

Figure 5.2: Behavior of cuttable food items in the simulated environment.

5.1 Novel Dataset

We create a replication of the kitchen depicted in the MPII Cooking 2 dataset using the Unity game engine[1], including all food ingredients, tools, and furniture from the original dataset. Our VACE dataset [170], [171] can be publicly downloaded[2]. In this VR environment, users can manipulate all furniture, tools, and ingredients by using a VR controller and observing through a VR headset. All graspable objects are rigid and behave in a physically plausible way with respect to gravity and friction. Actions such as *picking up*, *placing down*, *pushing*, *dropping*, *throwing*, and *stacking* objects are possible. Food items can be repeatedly *cut* into smaller pieces (without predefined section planes) using the usual cutting tools like knives, peelers, and graters. One difference to the VRKitchen dataset [156] is that there currently are just a few object state changes implemented. This includes opening and closing cupboards, the fridge, containers such as boxes with lids, pots with pot tops, and drawers, as well as opening and closing a water faucet and turning the different hot plates of a gas stove on and off. Even with missing object state changes, a reasoning agent can still imply that if an ingredient is placed in a pan on an active hotplate, then the object will be cooked, fried, or something alike.

At submission, our dataset contains 22 reenacted MPII Cooking 2 samples and new recipes performed by one user (10 × cut cucumber, 4 × cut bread, 4 × prepare salad, 4 × prepare sandwich). The goal is to have different recipes that share some steps of the preparation process, i.e., different goals with congruent plan prefixes. There are also multiple samples per dish such that there are different concrete goal states per goal (e.g., which kind of tableware is used) and ways to reach a specific goal state (e.g., whether to boil water in a kettle or in a pot before pouring it into another vessel). Variations currently include *with/without washing of the ingredients*, *with/without tidying up after preparation*, *with knife/with grater*, order variations (*get tools first/get food items first*), salad variations (*with/without additional spices*, *with/without stirring after seasoning*, *with/without pouring the salad into another bowl after stirring*), and sandwich variations (*bun/toast*).

As a concrete example, the preparation of a salad can be accomplished in various ways: The actor can choose which tools to use (e.g., one or multiple knives, forks, ladles, or the grater), whether to cut ingredients on the kitchen table surface, a cutting board, or a plate, whether to wash the vegetables before cutting them in the sink, which bowls to use, and how to season the salad, among others. Samples can also include actions like tidying up the kitchen after the food preparation, such as washing the used tools

[1] https://unity.com
[2] https://sites.google.com/view/vacedataset, doi:10.48436/r5d7q-bdn48, doi:10.48436/9y2x1-q4n71

Figure 5.3: A view inside the cupboards of the VACE scenario.

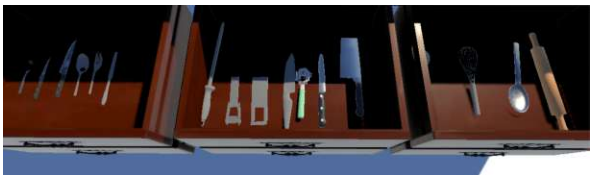

Figure 5.4: A view inside the drawers of the VACE scenario.

and storing them. The ordering of these steps can be changed, e.g., in which order to get the tools, containers, and food items from their respective storage places.

The kitchen environment is equipped with ∼40 types of kitchen tools, tableware, containers, ∼50 different types of food, as well as furniture like cupboards, tables, sink, stove, and fridge (Figure 5.3, Figure 5.4, and Table A.7.6 for an overview of the object types in the dataset). Similar to the MPII Cooking 2 dataset, there are two different types of samples: complex dishes consisting of multiple preparation steps, e.g., making a pizza, or a salad, and simple dishes, e.g., slicing avocados, onions, and the like.

5.1.1 Data Annotation

To make the generated dataset useful for as many research projects as possible, one sample consists of the following data, recorded at 30 Hz, see Fig. 5.1:

- RGB camera (1600 × 1200 px),
- depth camera (1600 × 1200 px),
- instance and class segmentation (640 × 480 px), pixel-wise color coding per object,
- object-wise bounding boxes in screen coordinate system,
- 3D human pose joint position of 20 joints. Head, hands, chest (via VR tracker), other joints from VR avatar calculated through inverse kinematics,
- 3D object poses and orientation,
- ground truth object predicates (*grasping, on, in, cutting, pushing*).

The ground truth object predicate annotation of the first and last frame of a sample can be used to generate an initial and final state of the sample for classical planning approaches. Bounding box, pose, orientation data, and logical predicates are stored per frame per object or per event (e.g., *cutting*) in the JSON format.

In Fig. 5.5 an example of a preparation task from the perspective of the human is shown. The recipe for a slice of bread with cheese and produce is based on the example in [156]. The steps include getting all necessary kitchen utensils, i.e., a knife and cutting board from a drawer and a plate from the cupboard. Then all the food utensils are gathered, namely cheese, cucumber, and tomato from the fridge, as well as a piece of bread from the cupboard. Next, the cook washes the cucumber and tomato in the faucet, needing to turn the faucet on and off. Then the cheese is cut, and a piece of it is put onto the bread. The bread is put in the oven to simulate baking the cheese (without a visual change of the piece of cheese, though). Then the bread is taken out of the oven again and put back on the table. Next, the produce is cut and some pieces are put onto the bread. Lastly the completed bread is put onto a plate for presentation. None of the described actions is hard-coded, i.e., if an object needs to be retrieved from the cupboard, one needs to grip the handle of the cupboard door, pull it open, reach inside, and grasp the desired object.

Fig. 5.2 shows in detail the behavior of objects being cut. Only food items can be cut, and only designated tools allow for cutting, namely knives, peelers, and graters. As soon as a cutting tool collides with a cuttable food item, the ingredient is sliced into two parts along the cutting plane of the tool from the point of contact, e.g., the blade.

5.2 Open Source and Interactive Simulator

We release this Unity project as an open-source tool[3] so that researchers can adapt the simulator to suit their needs, such as additional annotation, new objects, settings, interfaces, and the like. With the current recording procedure, one can record additional samples to add to the main dataset or create a new dataset. The workflow is split into a recording phase and a rendering phase. In the editor, the starting setup can be specified, like the arrangement of objects and furniture. Users can control the human avatar with a HTC VIVE headset[4]. For improved realism, it is recommended to additionally wear a HTC VIVE tracker[5] on the chest, which allows independent movement of the avatar's head relative to the body. The body of the avatar is animated through an inverse kinematic system in order to generate realistic arm, leg, and torso behavior.

For guidance during the recording of a task, users can choose any recipe of the MPII Cooking 2 dataset and then sequentially display each single cooking step of a chosen recipe in a head-up display and on a wall screen in the environment. The recipes are stored in a folder of the project and provided by the original MPII Cooking 2 dataset. For guidance in other recipes, users can provide their own recipes following the used encoding. Through the replay capability of pre-recorded samples in the simulator,

[3] https://github.com/michaelkoller/vacesimulator
[4] https://www.vive.com/eu/product/vive-pro-full-kit/
[5] https://www.vive.com/eu/accessory/tracker3/

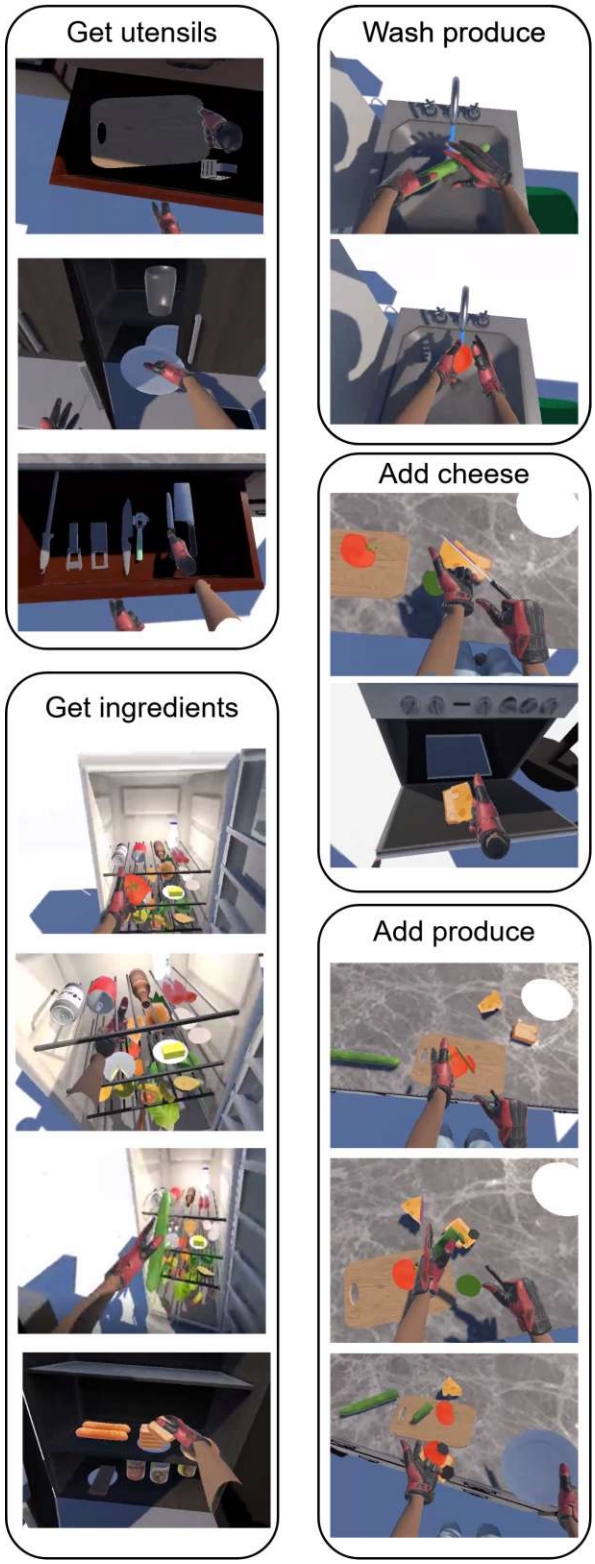

Figure 5.5: Steps in preparing slice of bread with baked cheese and produce from the ego perspective.

it is possible to re-annotate a sample or extend the previous annotation.

5.3 Conclusion and Future Work

In this paper, we presented VACE, a new VR simulator and synthetic dataset for recording object interaction-heavy cooking tasks in a richly furnished, interactable virtual kitchen. The dataset contains 22 samples with variations of 4 different recipes. Each sample is composed of RGB, depth, and segmentation mask images, as well as pose, orientation, bounding box data as well as logical predicates. Recording samples with the simulator does not require users to label samples per hand.

on the dataset and compare the results to other available datasets [172]. We believe that our simulator will allow researchers to create complex dynamic sequences with a fine-grained level of annotation suitable to their research. With proper spatial and semantic annotation of objects and atomic predicates, higher-level planning and recognition approaches can be explored more efficiently. The possibility to re-annotate existing samples (as was done with the VACE Ego Perspective Dataset, doi:10.48436/9y2x1-q4n71) and extend the dataset provides additional utility.

Limitations of our work include that users have to be trained in order to reliably grasp and stack objects. Users have to learn the kitchen layout, and how to handle the Unity software and VR hardware. At submission, only HTC Vive with one additional tracker is supported. This can be expanded in the future, and users are asked to contact us with feature requests. Developers must be familiar with Unity and C#.

Chapter 6

Preferences and Biases

On the previous time scale of up to several minutes, the robot's goal in joint attention scenarios was to determine the current intention of the observed human by estimating the task goal and plan. However, in a larger time scale of hours, days, and more, a long-term deployed robot will have many opportunities to observe the choice behavior of the human in a larger context. This means the robot can learn the preferences of the human, given a problem, its possible actions, and their outcomes.

In such decision situations, it has been shown that humans do not act purely rationally but display systematic irrational biases. Such cognitive biases include hyperbolic discounting [173], several availability biases [174], or the need to reduce cognitive dissonance [175], among many others. One well-documented cognitive bias is described in the Cumulated Prospect Theory (CPT) [29], [176], and it describes how humans systematically make suboptimal decisions in repeated games due to a cognitive distortion of probabilities and reward values regarding a specific anchor point.

In the previous Chapter, the plan recognition algorithm necessarily incorporates an estimation of goal priors [112]. If no other information is available, these can be modeled as a uniform distribution over the goal set. However, long-term deployment of a robot is well suited to transform these priors over time when the robot observes the actual preferences of a human in a given task domain. This human choice behavior could be seen as ground truth, but here we argue that this prior is the result of a human acting according to their biases. Knowing how to counteract a known bias would lead the prior distribution closer to the real distribution, which then improves the plan recognition process (Figure 6.1). That means that the prior calculated from human observation does not actually represent the preferences of the human, and we should look for a way to improve the priors.

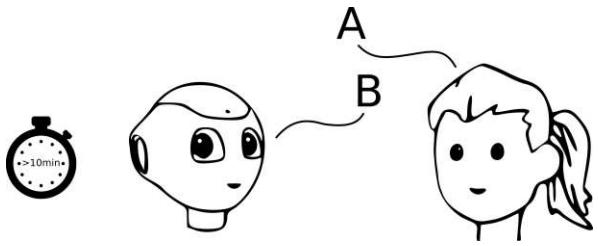

Figure 6.1: In the time resolution of ten minutes and above, the robot can update the belief about the preferences and biases of the human between repeated interactions. During interactions, it can use the updated estimation to propose more rewarding actions.

Recently, a formalism has been developed in which artificial agents are deployed to observe and possibly correct faulty human choices, called the assistive multi-armed bandit problem [28]. Specifically, an autonomous system observes and intercepts the repeated actions of a human, estimates the actual utility of the different actions, and potentially chooses a different action than the human to improve the overall return. Previous work, however, dealt only with a cognitive bias in a single interaction [177], i.e., not a repeated game, or with abstractions of human behavior that acted noisily rational [28], [178]. A noisily rational agent does not act according to a bias that shifts the perceived mean of a reward. Empirically observed human biases, however, transform the true mean reward of an action, making the problem more complex than filtering out the noise. Previous approaches would thus only learn to replicate the bias.

In this chapter, we present a framework that allows studying an abstract human policy that acts according to the cognitive bias defined in the CPT in repeated interactions. We also derive an algorithm that leverages knowledge about the risk-averse human model to correct the human bias in a human-robot team. We show in a multi-armed bandit experiment that the same robot agent improves the performance over a human agent acting alone in a scenario where risk aversion is suboptimal and that the performance of the human-robot team does not degrade below the performance of a rational agent in a scenario where risk aversion actually leads to a higher reward.

This result is an initial step for a long-term human-robot interaction style, where the robot can improve the expected return of the human for a given goal set in a task, even if the human deviates from a noisily rational or otherwise informative choice behavior.

Humans frequently find themselves playing what is known as multi-armed bandit (MAB) games (Figure 6.2, left), e.g., when choosing music to listen to, ordering food, making financial decisions, or committing to a plan to fulfill a task. In such settings, an actor repeatedly chooses an action (or pulls an arm, in analogy with the one-armed bandit gambling machine) without complete knowledge about the associated reward distribution of each action. During an episode, there is a trade-off between choosing what previously yielded the best results (exploitation) and choosing other actions to improve the estimate of the mean reward (exploration).

Social robots will also often face team situations, in which case they observe a human partner who themselves is not quite sure what they want or what the robot understands as both learn about the different outcomes. However, when an autonomous system in

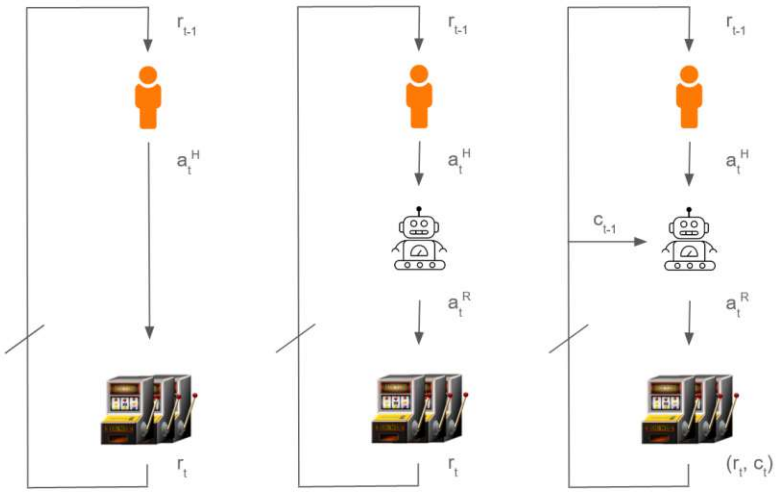

Figure 6.2: r_t, reward at time step t. a_t^H / a_t^R, action of the human / the robot at time step t. c_t, the reward class at time step t. Left: The standard MAB scenario, where a human repeatedly chooses among a group of actions to maximize the long-term reward. Middle: The assistive MAB scenario, where a robot observes and intercepts the action of the human and potentially chooses a different action. Right: The assistive MAB scenario with observable reward classes, where the robot observes the reward class additionally to the chosen human action (but not the actual reward value).

general or a robot specifically teams up with a human in such a situation, the goal is still to improve the overall utility for the human. Examples include under-specified commands such as "Set the table." or "Fetch me some food for lunch.". One way to model such a collaborative setting is the *assistive multi-armed bandit* [28], [178] (Figure 6.2, middle). Over multiple rounds, the human chooses an action of a multi-armed bandit with initially unknown rewards. Each round, the choice of the human is intercepted by a robot. The robot can then choose to perform the chosen action or choose another one. The human observes the robot's chosen action and the resulting reward, which influences their next choice. Thus, the robot must infer the reward only by the actions of the human.

With regard to HRI settings, humans are known to display systematic biases in decision-making scenarios like time-inconsistent planning due to hyperbolic discounting [179], false beliefs due to overly trusting attitudes [180], or risk-averse behavior in uncertain situations [177]. However, the predominant human model in previous work is that of a noisily rational agent, where sub-optimal decisions are modeled as noise. Furthermore, in the described scenario, the human also learns about the rewards and thus behaves in a non-stationary way, which is not modeled by a noisily rational agent. For the first problem, mathematical formulations to capture the bias exist [29]. For the second problem, there are multiple classes of bandit algorithms that can achieve good performance, but the problem of a risk-averse biased player has not yet been explored in MAB problems.

In behavioral economics, biases due to uncertain outcomes in risky situations are described by the (Cumulative) Prospect Theory [29], [176], which states that humans prefer certain over uncertain outcomes, weigh differences between small values stronger than the same difference between larger values, and perceive losses more strongly than gains. This formalism can explain phenomena such as gambling and insurance and has strong empirical evidence [181].

It is thus worthwhile to include these biases not only in single decision scenarios but also in repeated games with learning agents. This is achieved by transforming the statistics of probabilities and rewards that are associated with an arm of the bandit over time. As the robot cannot make inferences about the variance of rewards by only observing the action taken by the human each turn, the problem setting in this work is expanded in such a way that the robot is allowed to observe the human choice and the resulting reward class, without knowing the biased utility for the human or the true unbiased utility.

In this chapter, we (1) extend the human model for single decision games from [177] to a multi-armed bandit setting and study the properties of robot assistants who help humans that are modeled through different biased MAB learning algorithms. Further, we (2) expand the assistive MAB setting to observable reward classes without knowledge of the utility (Figure 6.2, right). This is then a preference learning problem not for the available choices but for the observable rewards. In a first approach, we (3) formulate an algorithm that has access to the human model - this can be understood as a ToM approach - that tries to assist the human by fitting reward values to make the history of human choices explainable under the known biases, then applying the inverse transformation to these estimations and making a choice based on these estimated unbiased values.

We (4) present an initial exploration of the performance of the risk-averse biased upper confidence bound (RAB UCB) human policy and the human-robot team consisting of a RAB UCB agent and a robot assistant. We show that a robot assistant will improve the overall return in situations where risk aversion is detrimental and that the team does not over-correct with respect to a baseline in situations where risk aversion is favorable.

This chapter is organized as follows. Chapter 6.1 covers other related work. In Chapter 6.2 the formalism for the assistive multi-armed bandit problem with observable reward classes, the risk aversion bias postulated by CPT, and the risk-averse biased upper confidence bound human policy are defined. In Chapter 6.3, an algorithm for a robot assistant is proposed. In Chapter 6.4, experimental results are presented. Chapter 6.5 concludes the chapter with a discussion. This chapter's content is based on previously published work in [182], [183].

6.1 Related Work

6.1.1 Risk-Averse Multi-Armed Bandit

In the standard MAB problem, the goal is to maximize the expected return but other variations attempt to also minimize the risk of incurring large losses [184]–[187].

Algorithms that solve this risk-averse MAB problem are different from the human model in this work, as they explicitly optimize a goal that includes some measure of risk. The risk-averse human model still only tries to maximize the expected return, but perceives rewards and probabilities through the risk-averse transformation by the Cumulative Prospect Theory transformation.

6.1.2 Other Human-Robot Team Settings

[178] explores the benefit of robots that disobey human orders if they have a sufficiently accurate estimation of the reward parameters. The agents perform in a POMDP with a set of featurized world states and static reward parameters. The reward is a linear combination of both. The human is modeled as a noisily rational agent with knowledge of the true reward parameters.

[180] describes semi-cooperative games in a POMDP setting between a human and a robot, where the robot has different assumptions about the degree of trust that a human has towards the goals of the robot (i.e., whether the robot tries to help them or not). It turns out that a robot with such a human model can in some cases decrease its own cost and induce the human to do more of the shared workload. The robot solves a POMDP in advance of the game without any learning afterward.

In [188], the turn order in the human-robot team is actually reversed. The robot can choose to behave autonomously (including doing nothing) or delegate a decision to a human, who is believed to know the true reward parameters. In this setting, the robot can learn to strike a balance between autonomy and safety.

[179] addresses another known human bias, namely hyperbolic discounting, which can lead a human to prefer different outcomes as the reward payout approaches in time. They devise a planning algorithm that proposes paths to goals that avoid temptation for such a time-inconsistent agent.

6.2 Formalism

In this section, we gather the preliminaries that allow us to define the Assistive Multi-Armed Bandit Problem with Observable Reward Classes, and a human policy that behaves in a risk-averse biased way in repeated games. The assistive multi-armed bandit problem is expanded by allowing reward class observation because knowledge of the risk aversion bias in human players is only useful when being able to gather estimates of the variance of outcomes.

6.2.1 Multi-Armed Bandit

The *multi-armed bandit* (MAB) $\mathcal{M}$ examined here is defined by:

- M: the number of different rewards
- N: the number of arms
- u: a distribution over $\mathbb{R}$

- v: distribution over space of distributions over the sets of size M.

At the start of the game, M different rewards are sampled from u, forming the reward set $\mathcal{R} = \{r^{(1)}, ..., r^{(M)}\}$, then for each arm $i \in \{1, \ldots, N\}$, a distribution $v_i = \{p_i^{(1)}, \ldots, p_i^{(M)}\}$ over $\mathcal{R}$ is sampled according to v. In each round t, the agent chooses an arm $a_t \in \{1, \ldots, N\}$ where a reward r_t is sampled from the respective distribution q_{a_t} over $\mathcal{R}$. The mean μ_i, $i \in \{1, \ldots, N\}$ per arm, called *consequence* in [177], is defined as

$$\mu_i = \sum_{j \in \{1,\ldots,M\}} p_i^{(j)} \cdot r_i^{(j)}. \tag{6.1}$$

$T_t(i)$ represents the number of arm pulls of the i-th arm up to time t.

6.2.2 Inverse Multi-Armed Bandit with Observable Reward Classes

In [28], the inverse multi-armed bandit problem is defined as a passive inference problem with a MAB $\mathcal{M}$ and a human H who employs a bandit strategy that maps histories of past actions and rewards to distributions over arm indices $H_t : h_1 \times r_1 \times \cdots \times h_{t-1} \times r_{t-1} \to \Pi(N)$. In the inverse multi-armed bandit problem with observable reward classes, the observing agent R knows the amount of different rewards M, but not their numeric values. The reward classes are labeled $c = \{c^{(1)}, \ldots, c^{(M)}\}$. The robot's goal is to infer their numeric value in order to recover $\Pi(N)$, while observing the choice of the human and the reward class each turn. This can be understood with the example of a slot machine with multiple arms, where each arm has a different distribution of the rewards 0, 5, and 10 gold coins. However, the observer R only sees them as rewards A, B, C and has to estimate the value of these rewards.

6.2.3 Assistive Multi-Armed Bandit with Observable Reward Classes

This is the active problem, where the joint system of a human and a robot $H \circ R$ play a MAB $\mathcal{M}$. Each round, the human player H chooses an arm a_t^H according to a given bandit strategy and the history of arm pulls and rewards $H(a_1^R, r_1, \ldots, a_{t-1}^R, r_{t-1}) \in [1, \ldots, N]$. But before the action is performed, the robot R intercepts and observes the chosen action a_t^H, and can then perform the human's choice or pick another action a_t^R, which is actually performed. The human then observes this action and the associated reward, while the robot observes the human's choice and resulting reward class $R(a_1^H, c_1, \ldots, a_{t-1}^H, c_{t-1}) \in [1, \ldots, N]$.

6.2.4 UCB Family of Bandit Algorithms

There are several algorithms that belong to the index-based family of bandit algorithms. They balance exploitation and exploration through the combination of a history of

previous pulls and rewards per arm $Q_t(a)$, $a \in [1, \ldots, N]$ (thus estimating their means) and an exploration bonus per arm (prefering arms that were less often pulled), among them Upper Confidence Bound (UCB), Bayes UCB and, Upper Credible Limit (UCL). The biases that are used in this work are applied to the history that an agent keeps while playing the MAB.

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbf{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbf{1}_{A_i=a}} \tag{6.2}$$

The arm with the maximum index each turn or a variation thereof with noise or softmax is chosen each round t, with $c \in \mathbb{R}^+$ balancing exploration and exploitation.

$$A_t = \text{argmax}_{a \in \{1,\ldots,N\}} Q_t(a) + c\sqrt{\frac{\ln t}{T_t(a)}} \tag{6.3}$$

6.2.5 Risk-Averse Biased Upper Confidence Bound

This section motivates the expansion of the setting of the assistive multi-armed bandit by the observation of reward classes. An observer who wants to estimate the utility of the reward classes through actions that are biased by the following transformation needs to be able to track the variance of outcomes for each given arm, as the transformation is performed on the reward values and probabilities of each arm.

In previous work, there are cases where the human agent playing a MAB is rational, noisily rational (i.e., knowing the actual reward parameters per arm from the beginning), or a rational learning bandit policy that is greedy in limit of exploration. The estimated reward values (for learning agents) that motivate the choices of the agents are congruent with the actual reward values and cannot model sub-optimal behavior apart from being noisy. The Cumulative Prospect Theory, on the other hand, models a systematic bias that shifts the arm means. Biased agents will maximize the return under this perceived bias. Any observer that employs merely a maximum likelihood method for explaining the human actions will only learn the biased means and then be able to filter out noise.

The following section describes the reward and probability transformation first described in [29] and recently used in an HRI scenario [177] (see Fig. 6.3) and expands them for in MAB problems. The human model used in this work is a UCB policy that incorporates a risk-averse bias into the estimated means $\hat{\mu}_a$ per arm a, then uses the biased estimated arm means and an exploration bonus to determine which arm to choose each round.

When playing, the human agent keeps a frequency statistic for each arm, where $\mathcal{R}(t) = \{(r^{(1)}, p^{(1)}, \ldots, r^{(L)}, p^{(L)}\}$, $L \in [1, \ldots, M]$ describes all rewards $(r^{(i)})$ and how often they have occurred $(p^{(i)} = T_{t-1}(i)/(t-1))$ among all arm pulls up to round $t-1$. $\mathcal{R}_a(t)$ describes these statistics per arm a at time step t. From this statistic, the unbiased $\hat{\mu}_a$ (eq. 6.1) can be estimated. In noisily rational models, given a mean per arm, the probability the human choosing an arm a can be modeled by

$$P(a) = \frac{\exp(\theta \cdot \hat{\mu}_a)}{\sum_{a \in \mathcal{A}} \exp(\theta \cdot \hat{\mu}_a)}, \tag{6.4}$$

with the *rationality coefficient* $\theta \in [0, \infty]$, which controls how noisy the choice is. The bigger θ becomes, the more rational the agent acts, while $\theta = 0$ corresponds to uniform random choice.

To arrive at a human model that not only adds noise to the choice, but shifts the estimated means per arm, the biased model uses a set of transformations first. Each reward value is transformed according to

$$v(R) = \begin{cases} R^{\alpha} & \text{if } R \geq 0 \\ -\lambda(-R)^{\beta} & \text{if } R < 0, \end{cases} \tag{6.5}$$

with $\alpha, \beta \in [0,1]$, and $\lambda \in [0, \infty)$. The coefficients α and β present how differences among rewards are perceived. λ models how losses have more or less importance than gains. Probabilities are transformed according to

$$w^{+}(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}}, \qquad w^{-}(p) = \frac{p^{\delta}}{(p^{\delta} + (1-p)^{\delta})^{1/\delta}}, \tag{6.6}$$

with $\gamma, \delta \in [0,1]$. The probability of positive and negative rewards are weighted by w^{+} and w^{-}, respectively. This transformation models how high probabilities are underweighted while small probabilities are overweighted. The probabilities of 0 and 1 remain unchanged.

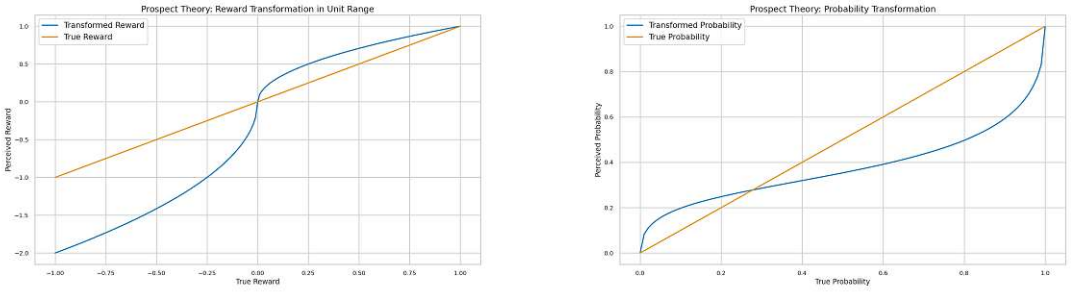

Figure 6.3: Transformations of the reward and probability in uncertain situations according to the Prospect Theory. Left: Reward transformation. Losses are weighted more strongly than gains. A difference between two big rewards is perceived as smaller than the same difference between two smaller rewards. $\alpha = \beta = 0.5$, $\lambda = 2$. Right: Probability transformation. Low probabilities are underweighted, while high probabilities are overweighted. 0 still maps to 0 and 1 still to 1. $\gamma = \delta = 0.5$.

Up to here, the transformation conforms to the Prospect Theory model, but the Cumulative Prospect Theory model proposes that weighting is applied more strongly to extreme events, i.e., high gains or losses, while Prospect Theory would weigh all events equally, no matter their magnitude. The following transformation captures this. Without loss of generality, the L rewards observed up to time $t-1$ are ordered in a decreasing order. Then the final probabilities weighted for each arm a are calculated according to

$$\pi(\mathcal{R}_a(t)) = (\pi^+(\mathcal{R}_a(t)), \pi^-(\mathcal{R}_a(t)), \tag{6.7}$$

$$\pi^+(\mathcal{R}_a(t)) = \Big(w^+(p_a^{(1)}), w^+(p_a^{(1)} + p_a^{(2)}) - w^+(p_a^{(1)}), \dots\Big), \tag{6.8}$$

$$\pi^-(\mathcal{R}_a(t)) = \Big(\dots, w^-(p_a^{(K)} + p^{(K-1)}) - w^-(p_a^{(K)}), w^-(p_a^{(K)})\Big). \tag{6.9}$$

Now the probabilities are normalized in order to sum up to 1.

$$\overline{\pi}_j(\mathcal{R}_a(t)) = \frac{\pi_j(\mathcal{R}_a(t))}{\sum_{i=1}^{K} \pi_i(\mathcal{R}_a(t))}, \forall j \in \{1, 2, \dots, K\} \tag{6.10}$$

$$\hat{\mu}_t^{CPT}(a) = \overline{\pi}_1(\mathcal{R}_a(t)) \cdot v\Big(r_a^{(1)}\Big) + \dots + \overline{\pi}_K(\mathcal{R}_a(t))) \cdot v\Big(r_a^{(K)}\Big) \tag{6.11}$$

The preference of the human is then described by

$$P_t(a) = \frac{\exp\Big(\theta \cdot \hat{\mu}_t^{CPT}(a)\Big)}{\sum_{a \in \mathcal{A}} \exp\Big(\theta \cdot \hat{\mu}_t^{CPT}(a)\Big)}, \tag{6.12}$$

and after adding the usual UCB exploration term the human finally chooses an arm according to

$$A_t = \text{argmax}_{a \in \{1,\dots,N\}} P_t(a) + c\sqrt{\frac{\ln t}{T_t(a)}}, \tag{6.13}$$

$$A_t = \text{argmax}_{a \in \{1,\dots,N\}} \hat{\mu}_t^{CPT}(a) + c\sqrt{\frac{\ln t}{T_t(a)}}, \tag{6.14}$$

with parameter c as exploration/exploitation trade-off factor.

6.3 Algorithm for Assistive Multi-Armed Bandits with Reward Class Observation

An initial approach to the problem is proposed by the following algorithm. Without a training phase, after a MAB $\mathcal{M}$ is instantiated with a given horizon T, the human policy H is instantiated with a set of parameters θ, α, β, λ, γ, δ, and c. The robot R policy is instantiated with a model of the human, i.e., it has an assumption about the human behavior and an initial value for the estimated rewards r_0. H has a descriptive statistic of previous rewards and their probabilities per arm, on which the CPT transformation is performed, and an arm pull counter for the exploration bonus. R keeps a history of human choices, robot choices, and reward class per time step $\mathcal{H} = [(a_t^H, a_t^R, c_t), \dots]$. The inverse reward transformation is

$$v^{-1}(R) = \begin{cases} R^{1/\alpha} & \text{if } R \geq 0 \\ -(-R/\lambda)^{1/\beta} & \text{if } R < 0. \end{cases} \tag{6.15}$$

Algorithm 2 Robot policy in H ∘ R Team

procedure CHOOSE ARM(t)
  create probability statistic $\boldsymbol{P}$ for $i \in [1, \ldots, t-1]$ from $\mathcal{H}$
  transform all $\boldsymbol{P}_i$ with eq. 6.6
  initialize $\hat{\boldsymbol{R}}$ with r_0
  minimize $\sum_{i=2}^{t-1} \text{argmax}_{a \in \mathcal{A}} \boldsymbol{P}_{i-1} \hat{\boldsymbol{R}} \neq a_i^H$
  inverse reward transformation on $\hat{\boldsymbol{R}}$ with eq. 6.15
  choose $a_t^R = \text{argmax} \boldsymbol{P}_{t-1} \hat{\boldsymbol{R}} + c\sqrt{\frac{\ln t}{T_t(a)}}$ ▷ State-action value + exploration bonus
  observe human choice a_t^H
  H and R observe reward r_t after a_t^R
  update $\mathcal{H}$

There is currently no method involved to address a potential exploration behavior of the human policy, which introduces a confounding error to the reward estimation. With increasing t, however, the biased learned arm means become the deciding factor in the human policy. After an initial learning phase, the error induced by the exploration behavior is therefore negligible. A detail that is yet to be examined is that the proposed robot algorithm assumes the Prospect Theory model, not the Cumulative Prospect Theory model. This choice was made due to the higher sensitivity of the inverse CPT transformation to reward fitting errors.

6.4 Experiments

The experiments seek to answer the following research questions:

- **RQ 7** Risky-better: Can the H-R team improve the performance (i.e., average return) of a risk-averse biased human policy in scenarios, where the risky option has a higher expected return?
- **RQ 8** Safe-better: Will the H-R team's performance deteriorate below the performance of the rational UCB policy when the lower variance option yields the higher expected return?

The experiments use a horizon $T = 300$ MAB averaged over $N = 300$ trials and compare a baseline UCB policy, a RAB UCB human policy, and a human-robot team. The human policy is instantiated with $\theta = 1$, $\alpha = \beta = 0.5$, $\lambda = 2$, $\gamma = \delta = 0.5$, and $c = \sqrt{2}$. The CPT parameters are roughly motivated by the international survey of Rieger et al. [189]. We picked a plausible above-average λ to model risk aversion. The robot policy has knowledge of these parameters. For each trial, the policies are exposed to the same sequence of rewards per arm. More specifically, before each trial, each arm is sampled T times, and the resulting rewards are saved in this order per arm. Thich allows a fair comparison between policies since each encounters the same series of rewards when exploring.

We fix two representative MABs (see. Fig. 6.4) with $N = 2$ arms and $M = 3$ different rewards for comparison: In the *Risky-better* MAB, a higher return can be achieved,

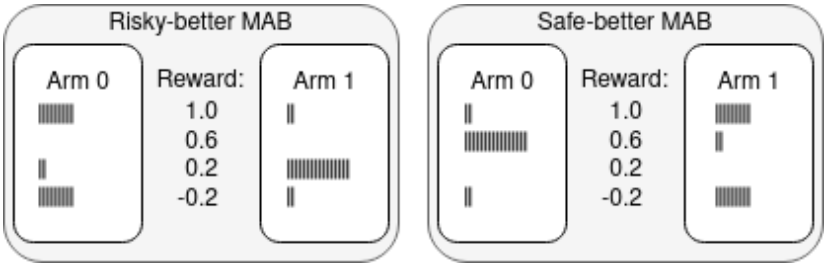

Figure 6.4: MABS used in the experiment. The bars under the arm name depict the probability of the reward outcome. Left: Risky-Better, Arm 0 ($\mu = 0.38, \sigma = 0.57$) has higher mean reward and higher std. than Arm 1 ($\mu = 0.24, \sigma = 0.28$). Right: Safe-Better, Arm 0 ($\mu = 0.56, \sigma = 0.28$) has higher mean reward and lower std. than Arm 1 ($\mu = 0.42, \sigma = 0.57$).

MAB	**Agent**	**avg. return**	**std.**
risky-better	UCB	102.25	11.57
	RAB UCB	82.05	9.13
	HR Team	87.55	11.45
safe-better	UCB	155.93	10.66
	RAB UCB	161.52	10.86
	HR Team	157.32	12.30

Table 6.1: Results for the comparison of two MABs ($N = 300$, $T = 300$) for three different agents.

when the arm with the more unsure and risky events is preferred. In the *safe-better* MAB, the opposite is true.

Table 6.1 shows the returns of the two MAB settings for each agent (see also Fig. 6.5 and Fig. 6.6). As expected, the H-R team moves the RAB UCB average return closer to the UCB agent. This means that in the risky-better MAB, the return improves, whereas the return deteriorates for the safe-better MAB. Closing the gap to the UCB agent, who acts (and explores) rationally, is a positive property of the human-robot team. It is unclear if there is a way to keep the overly optimistic behavior when an actor finds itself in an environment where risk aversion is indeed rewarded.

Performing a one-way ANOVA for the 3 agent types in the risky-better MAB reveals a significant difference: $F = 280.76$, $p < 0.001$. A Tukey HSD post-hoc comparison shows that μ_{UCB} is significantly higher than μ_{RABUCB}, $p < 0.001$ and μ_{HRT}, $p < 0.001$. Also, μ_{HRT} outperformed $\mu_{RABUCB}, p < 0.001$, therefore our assumption that the H-R team improves the performance (*RQ 7 risky-better*) is supported by the data.

Similarly, performing a one-way ANOVA for the three agent types in the safe-better MAB reveals a significant difference $F = 19.86$, $p < 0.001$. A Tukey HSD post-hoc comparison shows that μ_{RABUCB} outperformed μ_{UCB}, $p < 0.001$ and $\mu_{HRT}, p < 0.001$. However, there was no difference between μ_{HRT} and μ_{UCB}, $p = 0.291$. The data thus support research question *RQ 8 safe-better*.

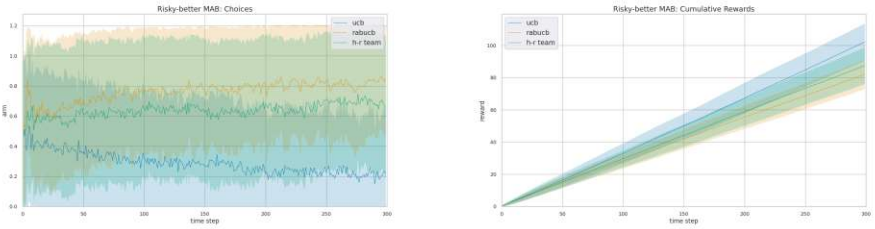

Figure 6.5: Results for MAB *risky-better* for UCB, RAB UCB and H-R Team ($N = 300$, $T = 300$). Left: Average choice over time. Lower is better, as arm 0 has a higher expected mean than arm 1. Right: Average cumulative reward over time. Higher reward is better.

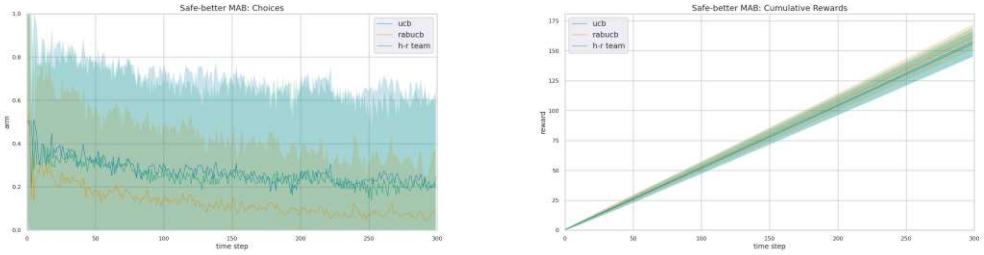

Figure 6.6: Results for MAB *safe-better* for UCB, RAB UCB and H-R Team ($N = 300$, $T = 300$). Left: Average choice over time. Lower is better, as arm 0 has a higher expected mean than arm 1. Right: Average cumulative reward over time. Higher reward is better.

6.5 Conclusion

In this chapter, we motivated the *Assistive Multi-Armed Bandit with Observable Reward Classes*. This expanded problem allows a direct observation of the variance of each arm in a MAB and, thus, a way to perform calculations involving variance. The original Assistive Multi-Armed Bandit problem essentially is transformed from a preference learning problem of arm choices into a preference learning of different discrete rewards. Continuous reward distributions can be handled by dividing the reward value range into a suitable amount of discrete bins. For the human-robot team, the previous single decision CPT framework is used to define the *Risk-Averse Biased UCB* policy, which captures empirical human biases. The robot assistant policy we proposed yields an initial approach to improve the overall return for the team by applying the CPT transformation to the observed probabilities each time step. Then it estimates the biased reward values from the history. Next, the inverse transformation is performed on the estimated biased rewards and the results are used to calculate the estimated unbiased arm means. It has been shown empirically that the Cumulative Prospect Theory models actual human behavior more accurately than the original Prospect Theory [181]. In this work, the human model acts according to the CPT, but the human model of the assistant policy assumes the human to act according to the Prospect

Theory. This design choice was made as the inverse transformation of CPT is more sensitive to reward estimation errors than the inverse transformation of PT.

6.5.1 Limitations

Further evaluations should compare different index-based policies with the CPT bias reward and probability transformation. This will shine a light on how different exploration-exploitation schemes impact the human-robot team. The influence of different CPT parameter ranges, as well as misscecified parameters in a human-robot team on the team performance should be experimentally measured. Future work can entail questions as to how robots can safely learn and leverage the human CPT parameters during a prolonged interaction.

6.5.2 Reactance

In the experiment, it became apparent that the human in the human-robot team chooses the suboptimal arm more often than the solo human policy. This is explainable by the index-based exploration bonus. As the human policy starts overestimating the mean of the suboptimal arm, the exploration bonus also increases, since the robot often chooses the other arm. Interestingly, the robot still can improve the overall return. A greedy biased human policy should turn out to be even more interpretable than the Risk-Averse Biased UCB. This algorithmic phenomenon lends itself to the very human interpretation of *reactance* [190]. Reactance describes an unpleasant motivational arousal in persons who are deprived of a previously available choice. Persons in such situations tend to exhibit a stronger preference for the lost option. Empirical research with human participants could shed light on whether a human enjoys being "helped" by a robot in such a way and how to design the interaction in a more pleasant way, taking in account the human arousal as an additional cost factor for example.

Chapter 7

Discussion

In the previous five chapters, we presented our work on joint attention-enabling mechanisms. We focused on the aspect of different time resolutions. This level of abstraction allows us to recognize different problems and provides the right frame to address different research questions. In this chapter, we integrate the findings and elaborate on how they relate in a joint action HRI scenario.

We proposed a specific high-level scenario: a socially competent domestic service robot that is long-term deployed in human living space. In [20], the authors demand *honest design* principles for robots with a recognizable head. This serves the purpose of not betraying the trust of interacting humans. The requirement imposes the restriction on the humanoid robot embodiment that the sensor capacity must coincide with that of a human, i.e., the robot can see in the direction that the head is pointing, and no more. The need for unobtrusive yet attentive interactions with humans with respect to a defined number of chore-like tasks characterizes this proposed deployment. Joint attention is, therefore, a relevant robotic capability for such scenarios.

In the highest time resolution (Figure 7.1), we were concerned with the gaze aversion ratio in an HRI scenario to determine how much time a robot can spend gazing at task-relevant locations and not at the interacting human. We found that the robot GAR influences the human GAR and interaction duration, but we found no significant differences in the consciously experienced interaction quality. Thus, we concluded that robot designers can take more freedom in robotic gaze design and diverge from human gaze timings in order to achieve additional interaction goals without deteriorating the conscious human interaction experience. This finding provides a useful justification for the relaxation of design requirements during an interaction and can be leveraged in the following problem, namely gaze sequence planning.

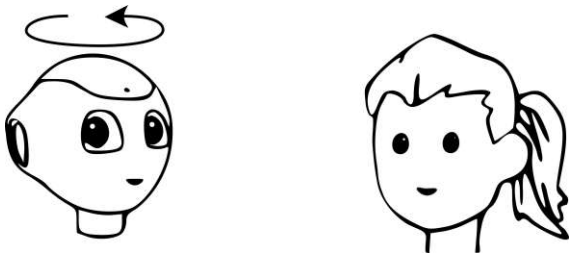

Figure 7.1: Gaze aversion in the time resolution of up to ten seconds.

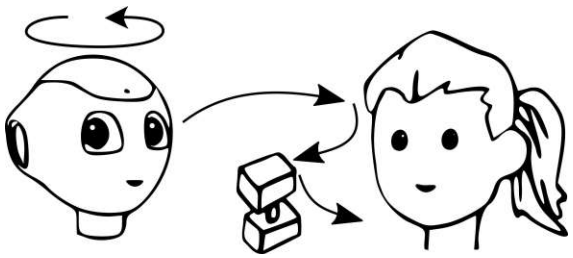

Figure 7.2: Gaze sequence planning in the time resolution of up to one minute.

In the second highest time resolution (Figure 7.2), we were concerned with how to chain different gaze targets together when a robot interacts with a human in a physical joint action task. The two competing purposes of the gaze target selection are task-relevant sensor alignment (i.e., pointing the head and, therefore, cameras at objects or locations in the environment) and socially interpretable joint attention signaling. This is achieved by learning a first-order MDP gaze transition function from human-human interactions. A state filter and an automated planner allow the robot to detect the current state of the task while the automated planner computes a plan to complete the task. Thus, the robot has an internal belief and intention, to use human terms. The gaze transition function is learned with respect to the dynamic task roles of the objects. When an object is next to be manipulated in order to achieve a task goal, it is declared the current active object. These roles are also defined for objects that are next or just previously in line and all other past and future objects. The learned gaze transitions are applied to the dynamic object roles (and not the objects themselves), the collaborator's hands, and head. The robot can combine sound task planning with gaze behavior from previous successful joint action tasks. The insight of our gaze aversion study simplifies the gaze sequence planning by only considering the chain of actual targets without the specific gaze dynamics, e.g., animation profiles and gaze durations.

The setting in the second-highest time resolution (Figure 7.3) operates under the assumption of only one possible goal. In general, however, there are multiple goals in the goal set. In a block stacking task, multiple block configurations could represent different goals, or different dishes could be prepared in a food preparation scenario. Thus, goal or plan recognition is the challenge in the third-highest time resolution of up to ten minutes. Previous work in plan recognition that relies on automated planning assumes that observation traces in the used planning language formalism are available. Robots, however, have only access to raw sensor data, in our example, video data. We derived an algorithm that leverages a common robot functionality, namely object detection, to generate observation trace candidates. This is done by leveraging the information inherent in physical object interactions in physical manipulation tasks. Thus, a robot can perform plan recognition without explicit action

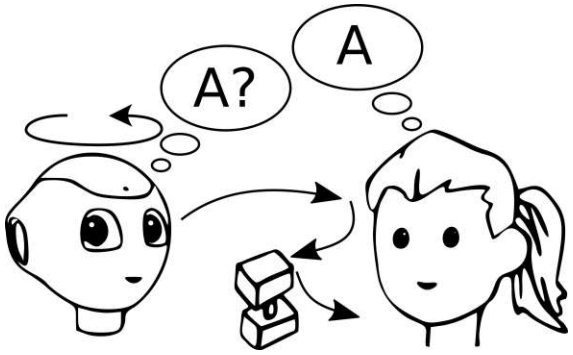

Figure 7.3: Plan recognition in the time resolution of up to ten minutes.

recognition. Using plan recognition and gaze sequence planning, a robot can now express its belief about the currently pursued goal of the human collaborator. More precisely, the robot observes the human actions, and computes a probability distribution over the goal set. The robot then adapts the joint attention-signaling gaze behavior according to the dynamic object roles of the most likely human plan.

We faced the difficulty of finding thoroughly annotated complex video samples of humans performing chore-like tasks with a long time horizon of up to several minutes. Therefore we created a VR simulation environment that allowed the recording of a synthetic dataset in a kitchen-like environment. Users can act in this environment through a virtual avatar using common VR equipment. The environment offers a multitude of interaction opportunities with respect to articulated furniture, tools, dishes, and food ingredients. Users can record samples of their food preparation tasks, which include multiple types of common video sensor data, ground truth pose annotations of all objects in the scene, and the logical scene state through first-order predicates. These complex samples with long time horizon and thorough annotation allow a standardized research effort for a multitude of research topics, e.g., how a robot should estimate not only the current goal of the observed human in a realistic setting but also when to intervene and with which action. For such cases, our environment provides long action sequences with realistic data samples and many different ways to reach a single goal. This offers flexibility for robotic interventions and also provides a challenging goal recognition task.

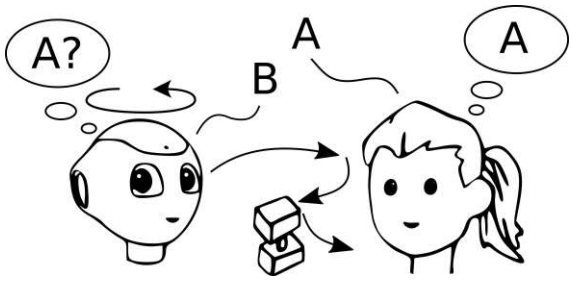

Figure 7.4: Preference and bias estimation in the time resolution of ten minutes and above.

A robot can estimate the intended goal of a human collaborator through the observed actions during a single interaction. However, during long-term deployment, a robot can also learn from the previous interaction with a collaborator and update its priors about the human's preferences. The time resolution of ten minutes and above (Figure 7.4)is the lowest time resolution in this thesis and is the frame of human preferences and biases. The assistive MAB, a recent human-robot team formalism, describes how robots can potentially improve the long-term reward of a suboptimal agent if they allow a robot to observe the choice and change it. However, these initial contributions abstract human behavior through noisy rational agents or other communicative suboptimal strategies. Human behavior, however, is shaped by non-communicative, systematically irrational biases, such as different kinds of risk awareness. We extended the setting of the assistive MAB by allowing the robot to observe reward classes. In each round, a robot can observe the human choice and the outcome of the action taken (which may not be the same action that the human chose). The reward classes in the assistive MAB can represent the goal set of the previously described time resolution. The robot knows that there are different outcomes. Each reward *class* corresponds to one goal in the

goal set. However, it does not know the reward *value* of the reward classes. The process of goal recognition in the previous time resolution is abstracted to recognizing a single action taken by the human, i.e., *choosing one goal*, and not considering how to reach the goal. The robot can ascribe different reward values to the different reward classes (i.e., goals) using the algorithm presented in Chapter 6. Thus, it can improve the plan recognition in the previous time resolution.

To summarize, the presented contributions interrelate and demonstrate how achieving gaze-related joint attention in HRI is a multifaceted problem:

1. We investigated how to design single gaze actions in social interactions.
2. We proposed a way to compute sequential gaze targets given one single goal.
3. We proposed an algorithm for plan recognition from object detection traces when there are multiple possible goals for a given task.
4. We published a simulator and dataset to study complex activities with a long time horizon.
5. We presented an algorithm that allows robots to correct human biases in an abstract reward maximization problem, which can be used to leverage learning opportunities during long-term deployment.

The research presented in the previous chapters was guided by the overarching research questions RQ 1-3 stated in Chapter 1.2.

7.1 RQ 1 - Gaze-Related Challenges in Joint Attention in Different Temporal Resolutions

RQ 1: Which HRI challenges can be identified by examining gaze-related joint attention through the lens of different temporal resolutions?

We provided conceptual approaches and solutions to problems of gaze-related joint attention in different time resolutions. We showed in Chapter 2 [42]–[44] that assumptions about human-inspired gaze timings in the second to second time resolution can be relaxed. In Chapter 3 [90] we showed how to use this relaxation in combination with HHI data to compute gaze sequences. In Chapter 4 [132], we leverage common robotic vision (object detection) and service robot capabilities (classical planning) to enhance the awareness of service robots and provide challenging data samples for robotic joint attention problems in Chapter 5 [169]. In Chapter 6 [182], [183], we contributed to the value alignment problem [191] in HRI when the robot possesses autonomy by expanding the assistive MAB formulation and introducing empirical systematic human biases in the human policy, and an algorithm to enhance human-robot team performance.

The core challenge of this research question stems from the multifaceted nature of the complex human phenomenon of joint attention, which manifests at different temporal resolutions. Psychological research results define processes of joint attention (e.g., RJA, IJA) and components of a cognitive architecture (e.g., structural model in [88], [89])

which can be used to partition the research problem into smaller pieces, as suggested in [21]. Psychological research results are most of the time descriptive, and thus it is the topic of technology-focused research to find technological correspondences to the single cognitive components.

Current technological systems model the different aspects of joint attention neither in its whole breadth nor depth. In order to conduct effective research, we need an exact description of the HRI research scenario with respect to the social and physical environment. In this thesis, we opted for the view of a domestic service robot as an intelligent tool that leverages the natural ability of humans to interpret social cues of non-human agents, such as reported in the apparent behavior of animated shapes [192], the media equation theory [59], and the intentional stance [193]. We argue that fluent task-related social behavior (i.e., gazing, pointing, object manipulation) is already a challenging research topic. In many instances in this thesis, it is difficult, if not impossible, to exclude emotional aspects of joint attention and joint action. Emotional interaction aspects are implicit in scenarios where the two agents have no conflicting goals (e.g., Chapter 3, How did the two agents agree on the goal in the first place?), and explicit in Chapter 2, where the effect of robot movement on the human user experience is measured, or Chapter 6, where the robotic intervention leads to behavior in the human policy which is akin to reactance in actual humans.

Eventually, the emotional aspect in joint action research cannot be neglected. However, currently robots still lack the mere task-related fluency in joint action scenarios which necessitates research that simplifies the emotional aspect of HRI in order to conduct effective research on task-oriented joint attention aspects. Similar to [21], we argue that as a preliminary step to joint action, future research should focus on the fluency in HRI as a user interface. Instead of relying on constant verbal feedback from the robot or the display of information on a screen, we argue that using the embodiment of the robot as the main communication modality in adequate scenarios is promising and will lead to high comfort, acceptance, and task-completion success. Thereby the focus should be put on achieving fluent interaction before achieving collaboration on shared goals between a human and a robot.

7.2 RQ 2 - Robot and Human Capabilities for Gaze-Related Joint Attention in Different Temporal Resolutions

RQ 2: What are the relevant robotic capabilities for gaze-related joint attention in the context of different temporal resolutions in comparison to humans?

The problem space of this research question is the mapping between empirically observed human capabilities and technological implementations on robotic systems. In Chapter 2 [42]–[44] , we leveraged the humanoid embodiment of a robot, face detection, speech, and animation principles. In Chapter 3 [90], we used face tracking, object detection, classical planning, and plan recognition together with empirical HHI data to recreate the gaze target distribution of a successfully cooperating human-human

dyad during a joint action task. In Chapter 4 [132], we leveraged classical planning and object detection to mimic human ToM capabilities in a robot. In Chapter 6 [182], [183], we used UCB policies and function approximation methods to model human learning from historic interaction data.

We found several correspondences between human faculties that enable joint attention and technological approaches to distinct research problems. Plan recognition in classical planning corresponds to ToM capabilities. Object and and face detection must be available as an initial step to process sensory data. However, a symbol grounding process must be in place not only the physical objects, but also perceived actions. All found correspondences are limited to certain scenarios and would not work in others. The main function of service robots is to manipulate physical objects and to navigate in an environment. Therefore, robotic capabilities must be robust for these domains and it must be clearly communicated to the user that the robotic capabilities are limited to these predefined areas. This can be done either by educating and informing users before and during contact with a service robot, but also by carefully designing the HRI itself. The importance of not overpromising robotic capabilities - both social and task-related - is apparent in the the anecdotal observations of failed social robots and the reported lack of long-term engagement with social robots [16].

Limiting the users' perceived intentionality [193] of a robot is interrelated with questions about how human the robotic behavior should appear (Chapter 2 and 3), and how autonomous the robot should be allowed to act (Chapter 6). Effective service robots will have to possess some degree of autonomy. The problem of value alignment [191] comes to the foreground when HRI leaves predefined research setups. Value alignment describes the problem of how to ensure that the programmed goal of a robot is congruent to the actual intention the programmer (and society in the larger context) intends to bestow on the robot.

Another core problem of realizing human-inspired behaviors on robots is that behavioral mappings for robots will never be totally congruent with HHI. Developing robotic capabilities that mimic human capabilities implies that human behavior must be modeled and re-imagined on a robot (unless one envisions an end-to-end implementation of HRI, but this approach currently seems infeasible due to its complexity). We found that is makes sense to apply a structural model of capabilities found in humans as an overarching design guide for robotic applications. However, for single systems in the structural model, we found it feasible to find translations of human capabilities to technological systems. For example, one aspect of ToM relevant for our scenario can be modeled through plan recognition. On a higher time resolution, an estimated goal and plan can be visualized by a gaze controller that produces head movement in the same statistical distribution of successful human-human joint interaction.

Another observation is that not only it is impossible for robots to employ unaltered human interaction behavior, but also that humans adapt their behavior when interacting with robots. It is likely that humans will try to make their actions legible to robots and thus differ from their usual HHI behavior.

Whenever a robot helps a human during a chore, the robot will employ some internal representation of the human in the computation of the next action. This representation will therefore have an effect on the human. In order to guarantee beneficial interactions

with robots for humans, the intrinsic value of human autonomy must be considered in collaborative scenarios. There must be an explicit weighting of the validity of human decisions, even when the robot is aware of human biases and is able to correct them.

7.3 RQ 3 - Existing Technologies for Gaze-Related Joint Attention Problems in Different Temporal Resolutions

RQ 3: Which existing technologies and approaches can be extended, modified, and used to solve gaze-related joint attention problems in different temporal resolutions?

In Chapter 2 [42]–[44], we were able to relax assumptions about the rigidity of human-inspired HRI approaches. We answered research questions by applying statistical tests on empirical data collected during conversational HRI. In Chapter 3 [90], we combined learned DTMCs with classical planning. In Chapter 4 [132], we combined ontological databases, classical plan recognition in the PDDL formalism, and MCTS. In Chapter 5 [169], we leveraged a game engine and VR interface to create a new dataset. In Chapter 6 [182], [183], we combined empirical psychological results about statistical parameters in the population with MAB, UCB, and function approximation algorithm.

The challenge in the technological conception of joint attention is to find approaches for the different facets of joint attention that either are separate to a degree where they do not impede each other or find technological approaches that solve multiple problems in the joint attention pipeline at once. The former variant applies to the separation of gaze dynamics (Chapter 2) and gaze targets (Chapter 3). The latter applies to the plan recognition approach (Chapter 4), where object detection and classical plan recognition are combined to not only arrive at an estimation of the current goal of the observed actor but also provide a plan to reach that goal. The robot can then use the plan to decide how to help the observed actor in reaching the goal. These methods are applicable in joint action interactions and complex tasks with long time horizon.

As it is often the case in robotics, for a working implementation in a real scenario, many different software components must be integrated. Each component fulfills a different role, e.g., object detection and pathfinding. For social scenarios, these considerations are valid as well. Social processes can be examined at different levels of abstraction. A couple of lower-level components are then combined to fulfill a higher function. At other times, a software component explicitly integrates data and output of lower-level components, as is the case with our gaze sequence controller or the plan recognition algorithm. As with all engineering tasks, different design approaches exist involving different lower-level frameworks. This thesis represents one of them.

In conclusion, the focus on different temporal resolutions for the contributions of this thesis allowed us to compile different but interrelating findings for the multifaceted joint attention HRI setting.

Chapter 8

Conclusion

The goal of this thesis was to enhance robotic joint attention capabilities during physical joint action tasks through design and computational insights. We offer a perspective on HRI challenges in different temporal resolutions and elaborate on their interrelation.

The main results of Chapter 2 reveal that on a behavioral level, a low gaze aversion ratio leads to shorter interaction durations and that human participants change their own gaze aversion ratio to mimic the robot. However, they do not copy the robotic gaze behavior strictly. Additionally, in the lowest gaze aversion setting, participants do not gaze back as much as expected, which indicates a user aversion to the robot gaze behavior. However, participants do not report different attitudes toward the robot for different gaze aversion ratios during the interaction. The urge of humans in conversational settings with a humanoid robot to adapt to the perceived gaze aversion ratio is stronger than the urge of intimacy regulation through gaze aversion, and high mutual gaze is not always a sign of high comfort, as suggested in previous work. This result can be used to justify deviating from human-inspired gaze parameters when it is necessary for specific robot behavior implementations.

In Chapter 3, we discuss human gaze behavior as an important modality for signaling, detecting, and monitoring joint attention processes. This is followed by an overview of joint attention implementations in HRI and commonly used artificial intelligence methods for planning and plan recognition. These methods are used to mimic the qualities of different components in psychological joint attention models in humans. In object manipulation tasks, the gaze behavior is not only used to gather information about the environment but also has a communicative role, as the interaction partner can interpret the gaze direction. The intended actions and beliefs about the current world state are communicated through the gaze. We argue that robotic gaze behavior, which humans easily interpret, will improve the interaction capability of a social robot. We investigate this claim in an already established HRI joint action benchmark scenario of collaboratively building a tower out of different blocks. To this end, we propose a stochastic gaze controller for joint action tasks and present the results of a pilot study.

In Chapter 4, we present an approach to estimate the goal of an observed actor from video data and provide a plan to achieve the goal. Since the resulting observation trace contains many noisy and redundant actions, a variant of the Monte Carlo Tree Search algorithm is used to construct a directed acyclic graph that compactly represents action

sequences. We focus on the context of object interactions and use video sequences with spatial object annotation traces in manipulation-heavy tasks suitable for robots. We show the results of a fragment of the MPII Cooking 2 dataset. We further contribute definitions and analysis of different metrics to estimate the goal with the highest posterior probability. Different metrics are defined to estimate the goal with the highest posterior probability.

In Chapter 5, we present the VACE dataset. Based on the MPII Cooking 2 dataset, it enables the recreation of recipes in a VR environment for a variety of meals and smaller activity sequences, such as cutting vegetables. For complex recipes, multiple samples are present, following different orderings of valid partially ordered plans. The dataset includes an RGB and depth camera view, bounding boxes, object segmentation masks, human joint poses, and object poses, as well as ground truth interaction data in the form of temporally labeled semantic predicates (holding, on, in, colliding, moving, cutting). In our effort to make the simulator accessible as an open-source tool, researchers are able to expand the setting and annotation to create additional data samples.

In Chapter 6, we expand the concept of the assistive MAB to improve the performance of human-robot teams where the human policy acts according to systematic irrational biases and not only noisily rational. The assistive multi-armed bandit setting is expanded by using observable reward classes but not their utility value. This allows deriving an algorithm that leverages knowledge about the risk-averse human model to correct human bias in a human-robot team. An evaluation indicates that arbitrary discrete reward functions can be handled.

In the discussion (Chapter 7), we provided considerations to integrating the different chapters, especially focusing on different temporal resolutions. We elaborated on the overarching challenges in joint attention research in HRI and answered the general research questions about challenges, mappings between human and robotic capabilities and combinations of technological approaches for gaze-related joint attention in HRI.

8.1 Limitations

Limitations specific to the individual contributions are presented at the end of the respective Chapters 2-6. In this section, we elaborate the overarching limitations of this thesis. One main limitation is that the findings of this thesis build upon each other conceptually, however, the individual contributions have not yet been integrated in one robotic system and empirically validated in user studies. This will include significant engineering efforts such as system integration for the many parallel mechanisms on a social robot.

During the thesis, we emphasized the different research challenges on the temporal resolution levels. For the topic of joint attention in HRI, this organization was productive. However, we do not claim that this is a definitive taxonomy for social processes. Additionally, the individual contributions focus on single aspects on the respective temporal resolution. For example, each represented research field has numerous results, which cannot easily be integrated in a single framework for HRI [63]. This consideration applies to diverse topics such as numerous different gaze parameters in conversational

settings and numerous different systematic biases in human decision-making.

Another limitation stems from the initial HRI setting assumption, namely that the interaction is always collaborative. This is an idealization, and it is far from being the typical case that human goals always align with the goals of a robot, no matter how well-intended the robotic programming is. Some examples include the effect of robot movements on the human user experience, miscalibrated expectations about the robot capabilities, agnostic or adversarial behavior of an observed actor in plan recognition, the degree of autonomy that leads a robot to overrule human decisions, and the value alignment problem in a broader scope.

As mentioned above, in this thesis we largely tried to exclude emotional aspects of joint attention HRI. However, the topic implicitly or explicitly appeared in several chapters. In Chapter 2, we tried to find a balance between the human user experience and design freedom in robotic gaze design. In Chapter 3 and 4, the robot excludes the emotional state of the collaborator and observed actor, although the emotional state influences human behavior. In Chapter 6, we noticed how the interaction between a human policy and a robot policy led to a behavior in the human policy akin to reactance, which is a cognitive state associated with strong emotions in humans.

8.2 Future Work

Avenues of future work in the respective research domains are mentioned in Chapters 2-6. The overarching direction of future work for this thesis is how humans interact with robots in different ways than with other humans. This applies to all aspects of this thesis, from gaze behavior to interactions where ToM capabilities are needed by both collaborators. Thereby the focus should be put on achieving fluent interaction before achieving collaboration on shared goals between a human and a robot.

Similar to [21], we argue that as a preliminary step to joint action, future research should focus on the fluency in HRI as a user interface. Instead of relying on constant verbal feedback of the robot or the display of information on a screen, we argue that using the embodiment of the robot as the main communication modality in adequate scenarios is promising and will lead to high comfort, acceptance, and task-completion success rates.

Moreover, humans apply different techniques when adapting their behavior towards robots, such as increasing the tone of their voice, or pushing a button repeatedly, among others. However, these attempts do not improve the interaction. Researching how these naturally occurring strategies arise as error correcting strategies can be used to generate further insights on fluent HRI.

Moving to more realistic scenarios and considering more aspects of joint attention in HRI scenarios simultaneously is a constant goal of HRI research and applies to the topics presented in this thesis as well. Specifically, increasing the complexity of the physical interaction setting and task domain is ongoing work and will provide more fluent HRI in the coming years.

Additionally, there is a need for more conceptual work on how we want to integrate robots into our everyday lives, knowing that it will change our own behavior. Without

design guidelines and an ethical vision of social robotics, future research is inapplicable at best, and detrimental to human society at worst. We estimate that the debate about domestic service robots with respect to their place on the spectrum between intelligent tools versus companions will intensify during the next years.

Appendix

A.1 Robot Script

We used the Pepper standard voice in English and the animations from the Pepper's movement library, provided by the NAOqi Python API.

Pepper Greeting:

- Say: "Hello!", Animation Slowly Offer Both Hands
- Say: "My name is Pepper!", Animation Both Hands Bump With Bump
- Say: "Please tell me about a movie you like! What is it about?", Animation Slowly Offer Both Hands

Pepper Farewell:

- Say: "Okay! Thanks for speaking to me.!", Animation Slowly Offer Both Hands
- Say: "I think, I now know enough about the movie.", Animation Both Hands Bump With Bump
- Say: "Thanks for your thoughts about it.", Animation Slowly Offer Both Hands
- Say: "Until next time! Bye!", Animation Happy

Pepper Parameters:
Speed values fall between 0 and 1, where 1 is specified as the maximum speed. Voice speed: 85, voice shaping: 110, gaze shift speed for pitch: 0.1, gaze shift speed for yaw: 0.2, gaze iteration duration: 10.0, gaze offset: 25, breath amplitude: 0.1, breath bpm: 10, blink duration: 0.05, speed postures: 0.3, speed head tracking: 0.1, head mov max range: 0.02, head mov min range: 0.01, speed speed random head movement: 0.3.

Pepper Gaze:
Gaze shifts are programmed to happen within 0.5s. When gazing away, the direction (left/right) is chosen randomly and a target head yaw is then ±25°. The joint position profiles are smoothly curved [1] and jerky movement is avoided. When directing the gaze

[1] https://developer.softbankrobotics.com/nao6/naoqi-developer-guide/naoqi-apis/naoqi-motion/almotion/joint-control, Case 3: Reactive Control

back from the averted state to the mutual gaze state, the last known angle position of the face of the human is targeted, until the human face is detected again. Then, the new position is targeted instead. A p-controller scheme is used to avoid jerky movements. While gazing at the human, pepper tracks the face of the human. Random head movement is applied, whenever the robot does not perform gaze shifts.

A.2 "Was that all?" - Robot Asked for More Information

	Attention		Comfort		Capability		Word count		Duration	
Question asked	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
No	3.62	0.953	3.87	0.768	3.28	0.912	332.	139.	185.	70.9
Yes	3.94	0.852	3.77	0.632	3.51	0.630	150.	65.4	93.2	36.1

Table A.2.1: Descriptive statistics between the two participant groups "robot asked for more: yes and no".

	Attention	Comfort	Capability	Word count	Duration
χ^2	3.284	1.548	0.536	52.678	53.702
p	0.069	0.213	0.464	<0.000	<0.000

Table A.2.2: Kruskal-Wallis tests for differences between the "robot asked for more: yes and no" conditions.

A.3 Human Gaze Data Evaluation

GAR	n	mean (ms)	median (ms)	SD (ms)
0.1	3783	372	219	372
0.3	4221	384	234	411
0.5	5799	406	250	453
0.7	4896	350	219	360
0.9	4830	330	219	310

Table A.3.3: Descriptive statistics for fixations per GAR condition. 0.1: Robot mostly stared at the human. 0.9: Robot mostly looked away from the human. Number of recorded fixations per condition (n), mean, median, and SD of fixation durations per condition.

GAR	mean	SD	median
0.1	0.65	0.19	0.65
0.3	0.69	0.2	0.69
0.5	0.65	0.2	0.65
0.7	0.54	0.22	0.55
0.9	0.51	0.17	0.46

Table A.3.4: Descriptive statistics for the normalized fixation duration on ROI *head* for all five GAR conditions.

A.3.1 Attitudinal Data Evaluation

	χ^2	p	η^2
attention	2.588	0.628	-0.015
comfort	4.685	0.321	0.007
capability	8.689	0.069·	0.051
artificial	2.578	0.630	-0.015
incompetent	8.107	0.087·	0.045
intelligent	6.026	0.197	0.022
sensible	11.533	0.021*	0.082

Table A.3.5: Kruskal-Wallis tests for the Likert scales *attention*, *comfort*, and *capability*, as well as the Likert items *artificial*, *incompetent*, *intelligent*, and *sensible*. Effect sizes: > 0.01 small effect, > 0.06 medium effect and > 0.14 (large effect). Significance level: '·': $p < 0.1$, '*': $p < 0.05$.

A.4 Participant Instructions

Written Experiment Introduction

Before signing a written consent form, participants read the following written introduction or the experimenter provided a verbal introduction following the written introduction. This is a short study on human-robot conversation. We want to know how people talk to a robot in certain situations to improve the human-robot dialogue. You help us a lot by participating. In this room you see the robot Pepper. Pepper speaks in English. You can talk in English, but if you feel uncomfortable, you can also talk in German. The robot will ask you about a movie you like. Just answer by describing one of the best movies you have watched recently in a couple of minutes and speak about it for about 2 minutes. The robot can see and hear you, just talk like you would normally talk to a friend. Your task is: Just answer by describing one of the best movies you have watched recently in about 2 minutes. Be aware that Pepper will not answer any questions. Pepper's task is: It will simply try to understand you. This study is done to gather data to improve human-robot dialogue, but it doesn't test you

in any way. Afterwards we will give you a questionnaire and ask you a few questions about your experience in person.

Verbal Experiment Introduction Before manually triggering the start of the experiment, the experimenter elaborated the following: So, this is the robot Pepper. It can hear and see you. Pepper will ask you to tell it about a movie. Your task is to just answer by describing one of your favorite movies or a movie you watched recently. Be aware that pepper will not answer any questions. It will simply try to understand you.

A.5 Questionnaire

Questionnaire Pepper's Movies

Participant#:
Condition:
Folder#:
Date, time:

Thank you for participating in our experiment. Please answer some questions for us!
(Questions about Pepper are on the back)

About you

Your age:

Your gender:

Your main profession:

Your previous experience with robots:

never	once	a few times	regularly
◯	◯	◯	◯

Your use of computers:

never	sometimes	daily
◯	◯	◯

Do you wear optical glasses for the experiment (apart from the eye tracker):

yes	no
◯	◯

Can you see the robot clearly:

yes	no
◯	◯

About Pepper:

1. Please rate your impression of **the way you were being attended** by the robot, based on the following statements:

please mark one oval per row

	Strongly disagree				Strongly agree	Don't know
Pepper made a **responsive** impression	☐	☐	☐	☐	☐	☐
Pepper made an **interactive** impression	☐	☐	☐	☐	☐	☐
Pepper made an **ignorant** impression	☐	☐	☐	☐	☐	☐
Pepper made an **unconscious** impression	☐	☐	☐	☐	☐	☐

2. Please rate your **feeling of comfort** with the robot, based on the following statements:

please mark one oval per row

	Strongly disagree				Strongly agree	Don't know
Pepper made a **creepy** impression	☐	☐	☐	☐	☐	☐
Pepper made me **feel nervous**	☐	☐	☐	☐	☐	☐
Pepper made a **warm** impression	☐	☐	☐	☐	☐	☐
Pepper was acting **pleasantly**	☐	☐	☐	☐	☐	☐

3. Please rate your impression of the robot's **interaction capabilities** based on the following statements:

please mark one oval per row

	Strongly disagree				Strongly agree	Don't know
Pepper's interactions were **artificial**	☐	☐	☐	☐	☐	☐
Pepper made an **incompetent** impression	☐	☐	☐	☐	☐	☐
Pepper was acting **intelligently**	☐	☐	☐	☐	☐	☐
Pepper was acting **sensibly**	☐	☐	☐	☐	☐	☐

A.6 Glossary

Optional glossary (Merriam Webster & Cambridge Dictionary)

responsive
quick to respond or react appropriately or sympathetically

interactive
Mutually or reciprocally active

ignorant
lacking knowledge or comprehension of the thing specified

unconscious
not knowing or perceiving | not aware

creepy
strange or unnatural and making you feel frightened

nervous
tending to produce agitation or nervousness | uneasy

warm
marked by or readily showing affection, gratitude, cordiality, or sympathy

pleasant
enjoyable, attractive, friendly, or easy to like

artificial
lacking in natural, lifelike qualities

incompetent
not having the ability to do something as it should be done

intelligent
able to learn and understand things easily

sensible
based on or acting on good judgment and practical ideas or understanding

A.7 VACE Object Types

utensils			
baking paper	detergent	ladle	small plate
big plate	dishsoap	large bowl	smartphone
blender	fork	mug	spatula
bowl	front peeler	pan	spoon
can opener	glass	paper towel roll	tea cup
cleaver	grater	pot	tea pot
coffee machine	handsoap	pot top	toaster
cup knife	rolling pin	towel	wine glass
cutting board	large knife	saucepan	tray
deep pan	knife sharpener	saucer	whisk
ingredients			
apple	carrot	kiwi	pepper
avocado	cheese	lemon	pineapple
baguette	chocolate bar	lime	potato
balsamic	cocoa	corn	salad
beer bottle	cucumber	milk	salt
beer can	dough	mushroom	sausage
blueberries	egg	mustard bottle	soy sauce
bread loaf	flour	olive oil	strawberry
broccoli	garlic	onion	sugar
bun	ground coffee	orange	toast
can	hot sauce	paprika	tomato
butter	ketchup bottle	pear	zucchini
can			
furniture			
counter	bin	drawer	stove
cabinet	sink	outlet	table
fridge	faucet	spice holder	
avatar			
human	left hand	right hand	

Table A.7.6: Overview of object types in the VR environment.

List of Figures

List of Tables

Bibliography

[1] A. Lambert, N. Norouzi, G. Bruder, and G. Welch, "A systematic review of ten years of research on human interaction with social robots," *International Journal of Human–Computer Interaction*, vol. 36, no. 19, pp. 1804–1817, 2020 (cit. on p. 1).

[2] W. Johal, "Research trends in social robots for learning," *Current Robotics Reports*, vol. 1, no. 3, pp. 75–83, 2020 (cit. on p. 1).

[3] C. Breazeal, K. Dautenhahn, and T. Kanda, "Social robotics," *Springer handbook of robotics*, pp. 1935–1972, 2016 (cit. on p. 1).

[4] M. I. Posner, C. R. Snyder, and B. J. Davidson, "Attention and the detection of signals," *Journal of Experimental Psychology: General*, vol. 109, no. 2, pp. 160–174, 1980 (cit. on p. 1).

[5] H. Pashler, "Processing stages in overlapping tasks: Evidence for a central bottleneck," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, no. 3, pp. 358–377, 1984 (cit. on p. 1).

[6] D. E. Broadbent, *Perception and communication*. London: Pergamon Press, 1958 (cit. on p. 1).

[7] R. Schulz, P. Kratzer, and M. Toussaint, "Preferred interaction styles for human-robot collaboration vary over tasks with different action types," *Frontiers in Neurorobotics*, vol. 12, 2018. [Online]. Available: `https://www.frontiersin.org/articles/10.3389/fnbot.2018.00036` (cit. on pp. 1, 52, 64).

[8] M. Scaife and J. S. Bruner, "The capacity for joint visual attention in the infant," *Nature*, vol. 253, no. 5489, pp. 265–266, 1975 (cit. on pp. 1, 42, 43).

[9] G. Knoblich, S. Butterfill, and N. Sebanz, "Psychological research on joint action: Theory and data," *Psychology of Learning and Motivation*, vol. 54, pp. 59–101, 2011 (cit. on p. 1).

[10] P. Mundy and L. Newell, "Attention, joint attention, and social cognition," *Current Directions in Psychological Science*, vol. 16, no. 5, pp. 269–274, 2007 (cit. on pp. 1, 42, 45).

[11] C.-M. Huang and A. L. Thomaz, "Joint attention in human-robot interaction," *2010 AAAI Fall Symposium Series*, 2010 (cit. on pp. 1, 47, 48).

[12] M. Tomasello, "Joint attention as social cognition," in *Joint attention: Its origins and role in development*, C. Moore and P. J. Dunham, Eds., Lawrence Erlbaum Associates, Inc., 1995, 103—130 (cit. on pp. 1, 43).

[13] A. Bauer, D. Wollherr, and M. Buss, "Human–robot collaboration: A survey," *International Journal of Humanoid Robotics*, vol. 5, no. 01, pp. 47–66, 2008 (cit. on pp. 1, 2).

[14] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind"?" *Journal of Cognition*, vol. 21, no. 1, pp. 37–46, 1985 (cit. on p. 2).

[15] C. Bartneck, *Why do all social robots fail in the market?* `https://www.human-robot-interaction.org/wp-content/uploads/2021/09/HRI-Podcast-Episode-015-Why-Do-All-Social-Robots-Fail-In-The-Market-Transcript.pdf`, Accessed: 2022-11-14, 2020 (cit. on p. 2).

[16] A. Weiss, A. Pillinger, and C. Tsiourti, "Merely a conventional 'diffusion' problem? on the adoption process of anki vector," in *2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2021, pp. 712–719 (cit. on pp. 2, 105).

[17] J. E. Laird, *The Soar cognitive architecture*. MIT Press, 2019 (cit. on p. 2).

[18] J. R. Anderson, *How can the human mind occur in the physical universe?* Oxford University Press, 2009 (cit. on p. 2).

[19] M. Brandão, "Normative roboticists: The visions and values of technical robotics papers," in *2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2021, pp. 671–677 (cit. on pp. 3, 4).

[20] M. E. Kaminski, M. Rueben, W. D. Smart, and C. M. Grimm, "Averting robot eyes," *Md. L. Rev.*, vol. 76, pp. 983–1025, 2016 (cit. on pp. 4, 11, 38, 100).

[21] N. C. Krämer, S. Eimler, A. Von Der Pütten, and S. Payr, "Theory of companions: What can theoretical models contribute to applications and understanding of human-robot interaction?" *Journal of Applied Artificial Intelligence*, vol. 25, no. 6, pp. 474–502, 2011 (cit. on pp. 5, 41, 43, 46, 104, 109).

[22] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: A review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 25–63, 2017 (cit. on pp. 5, 10, 11, 14, 55).

[23] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu, "Conversational gaze aversion for humanlike robots," in *2014 9th ACM/IEEE International Conference on Human-Robot interaction*, ACM, 2014, pp. 25–32 (cit. on pp. 5, 10, 11, 13, 29, 31, 36, 46, 57, 61).

[24] Y. Ban, X. Alameda-Pineda, F. Badeig, S. Ba, and R. Horaud, "Tracking a varying number of people with a visually-controlled robotic head," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2017, pp. 4144–4151 (cit. on pp. 6, 11).

[25] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction*, ACM, 2016, pp. 83–90 (cit. on p. 6).

[26] M. Ramırez and H. Geffner, "Plan recognition as planning," in *21st International Joint Conference on Artifical Intelligence*, Citeseer, 2009, pp. 1778–1783 (cit. on pp. 6, 7, 51, 66).

[27] S. Yang, Q. Gao, C. Liu, C. Xiong, S.-C. Zhu, and J. Chai, "Grounded semantic role labeling," in *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 149–159 (cit. on pp. 7, 81).

[28] L. Chan, D. Hadfield-Menell, S. Srinivasa, and A. Dragan, "The assistive multi-armed bandit," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction*, IEEE, 2019, pp. 354–363 (cit. on pp. 7, 88, 89, 92).

[29] A. Tversky and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and Uncertainty*, vol. 5, no. 4, pp. 297–323, 1992 (cit. on pp. 7, 87, 89, 90, 93).

[30] A. Gabouer, J. Oghalai, and H. Bortfeld, "Parental use of multimodal cues in the initiation of joint attention as a function of child hearing status," *Discourse Processes*, vol. 57, no. 5-6, pp. 491–506, 2020 (cit. on p. 9).

[31] M. F. Land, "Eye movements and the control of actions in everyday life," *Progress in Retinal and Eye Research*, vol. 25, no. 3, pp. 296–324, 2006 (cit. on p. 9).

[32] S. Devin, A. Clodic, and R. Alami, "About decisions during human-robot shared plan achievement: Who should act and how?" In *International Conference on Social Robotics*, Springer, 2017, pp. 453–463 (cit. on pp. 9, 41, 52, 53).

[33] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 333–356, 1988 (cit. on pp. 10, 41, 61).

[34] H. Kozima, M. P. Michalowski, and C. Nakagawa, "Keepon," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 3–18, 2009 (cit. on p. 10).

[35] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967 (cit. on p. 10).

[36] Y. Zhang, J. Beskow, and H. Kjellström, "Look but don't stare: Mutual gaze interaction in social robots," in *International Conference on Social Robotics*, Springer, 2017, pp. 556–566 (cit. on pp. 11, 15).

[37] A. Weiss, R. Bernhaupt, M. Lankes, and M. Tscheligi, "The usus evaluation framework for human-robot interaction," in *AISB2009: Symposium on New Frontiers in Human-Robot Interaction*, vol. 4, 2009, pp. 11–26 (cit. on p. 11).

[38] F. Delaunay, J. de Greeff, and T. Belpaeme, "A study of a retro-projected robotic face and its effectiveness for gaze reading by humans," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction*, 2010, pp. 39–44 (cit. on p. 11).

[39] U. J. Pfeiffer, B. Timmermans, G. Bente, K. Vogeley, and L. Schilbach, "A non-verbal turing test: Differentiating mind from machine in gaze-based social interaction," *PLOS ONE*, vol. 6, no. 11, pp. 1–12, Nov. 2011 (cit. on p. 12).

[40] N. Chen and P. J. Clarke, "Gaze-based assessments of vigilance and avoidance in social anxiety: A review," *Current Psychiatry Reports*, vol. 19, no. 9, pp. 1–9, 2017 (cit. on p. 12).

[41] N. Burra and D. Kerzel, "Meeting another's gaze shortens subjective time by capturing attention," *Cognition*, vol. 212, pp. 1–11, 2021 (cit. on p. 12).

[42] M. Koller, "Systematic variation of gaze timings and effects on the human level of comfort and feeling of being attended," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction*, IEEE, 2019, pp. 721–723 (cit. on pp. 12, 103, 104, 106).

[43] M. Koller, D. Bauer, J. de Pagter, G. Papagni, and M. Vincze, "A pilot study on determining the relation between gaze aversion and interaction experience," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction*, IEEE, 2019, pp. 644–645 (cit. on pp. 12, 17, 18, 103, 104, 106).

[44] M. Koller, A. Weiss, M. Hirschmanner, and M. Vincze, "Robotic gaze and human views - a systematic exploration of robotic gaze aversion and its effects on human behaviors and attitudes," *Frontiers in Robotics and AI*, vol. 10, 2023 (cit. on pp. 12, 103, 104, 106).

[45] N. J. Emery, "The eyes have it: The neuroethology, function and evolution of social gaze," *Neuroscience and Biobehavioral Reviews*, vol. 24, no. 6, pp. 581–604, 2000 (cit. on p. 13).

[46] V. Srinivasan and R. Murphy, "A survey of social gaze," in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction*, 2011, pp. 253–254 (cit. on p. 13).

[47] M. Argyle, *The psychology of Interpersonal Behaviour*. Penguin UK, 1994 (cit. on p. 14).

[48] M. Argyle, M. Cook, and D. Cramer, "Gaze and mutual gaze," *The British Journal of Psychiatry*, vol. 165, no. 6, pp. 848–850, 1994 (cit. on pp. 14, 32).

[49] H. Admoni, C. Bank, J. Tan, M. Toneva, and B. Scassellati, "Robot gaze does not reflexively cue human attention," in *Annual Meeting of the Cognitive Science Society*, vol. 33, 2011, pp. 1983–1988 (cit. on p. 14).

[50] J. Kennedy, P. Baxter, and T. Belpaeme, "Head pose estimation is an inadequate replacement for eye gaze in child-robot interaction," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction*, ACM, 2015, 35—36 (cit. on p. 14).

[51] G. Perugia, M. Paetzel-Prüsmann, M. Alanenpää, and G. Castellano, "I can see it in your eyes: Gaze as an implicit cue of uncanniness and task performance in repeated interactions with robots," *Frontiers in Robotics and AI*, vol. 8, 2021. [Online]. Available: `https://www.frontiersin.org/articles/10.3389/frobt.2021.645956` (cit. on pp. 14, 35).

[52] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "Affect recognition for interactive companions: Challenges and design in real world scenarios," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 89–98, 2010 (cit. on pp. 14, 35).

[53] F. Papadopoulos, D. Küster, L. J. Corrigan, A. Kappas, and G. Castellano, "Do relative positions and proxemics affect the engagement in a human-robot collaborative scenario?" *Interaction Studies*, vol. 17, no. 3, pp. 321–347, 2016 (cit. on pp. 14, 35).

[54] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1-2, pp. 140–164, 2005 (cit. on pp. 14, 35, 36).

[55] C. Yu, P. Schermerhorn, and M. Scheutz, "Adaptive eye gaze patterns in interactions with human and artificial agents," *ACM Transactions on Interactive Intelligent Systems*, vol. 1, no. 2, pp. 1–25, 2012 (cit. on pp. 15, 27).

[56] M. Argyle and J. Dean, "Eye-contact, distance and affiliation," *Sociometry*, pp. 289–304, 1965 (cit. on pp. 15, 35).

[57] P. Baxter, J. Kennedy, A.-L. Vollmer, J. de Greeff, and T. Belpaeme, "Tracking gaze over time in hri as a proxy for engagement and attribution of social agency," in *2014 9th ACM/IEEE International Conference on Human-Robot Interaction*, 2014, pp. 126–127 (cit. on pp. 15, 24, 35).

[58] P. K. Jokinen, "Conversational gaze modelling in first encounter robot dialogues," in *11th International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA), 2018, pp. 1–6 (cit. on pp. 15, 35).

[59] B. Reeves and C. Nass, "The media equation: How people treat computers, television, and new media like real people," *Cambridge, UK*, vol. 10, pp. 19–36, 1996 (cit. on pp. 15, 104).

[60] K. Ijuin and K. Jokinen, "Exploring gaze behaviour and perceived personality traits," in *2020 15th ACM/IEEE International Conference on Human-Computer Interaction*, Springer, 2020, pp. 504–512 (cit. on p. 15).

[61] Y. Sabyruly, F. Broz, I. Keller, and K. S. Lohan, "Gaze and attention during an hri storytelling task," in *2015 AAAI Fall Symposium Series*, 2015, pp. 1–4 (cit. on p. 15).

[62] F. Broz, H. Lehmann, C. L. Nehaniv, and K. Dautenhahn, "Mutual gaze, personality, and familiarity: Dual eye-tracking during conversation," in *2012 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2012, pp. 858–864 (cit. on pp. 15, 27).

[63] T. Belpaeme, "Advice to new human-robot interaction researchers," in *Human-Robot Interaction: Evaluation Methods and Their Standardization*, C. Jost, B. Le Pévédic, T. Belpaeme, *et al.*, Eds. Springer International Publishing, 2020, pp. 355–369 (cit. on pp. 17, 108).

[64] O. Johnston and F. Thomas, *The illusion of life: Disney animation*. Disney Editions New York, 1981 (cit. on p. 17).

[65] T. Ribeiro and A. Paiva, "Nutty-based robot animation–principles and practices," *arXiv preprint arXiv:1904.02898*, 2019 (cit. on p. 17).

[66] R. C. R. Mota, D. J. Rea, A. Le Tran, J. E. Young, E. Sharlin, and M. C. Sousa, "Playing the 'trust game'with robots: Social strategies and experiences," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2016, pp. 519–524 (cit. on p. 17).

[67] L. D. Riek, "Wizard of oz studies in hri: A systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 119–136, 2012 (cit. on pp. 18, 37).

[68] N. M. Fraser and G. N. Gilbert, "Simulating speech systems," *Computer Speech and Language*, vol. 5, no. 1, pp. 81–99, 1991 (cit. on p. 18).

[69] M. Hirschmanner, S. Gross, S. Zafari, B. Krenn, F. Neubarth, and M. Vincze, "Investigating transparency methods in a robot word-learning system and their effects on human teaching behaviors," in *2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2021, pp. 175–182 (cit. on p. 18).

[70] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009 (cit. on p. 19).

[71] M. Joosse, A. Sardar, M. Lohse, and V. Evers, "Behave-ii: The revised set of measures to assess users' attitudinal and behavioral responses to a social robot," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 379–388, 2013 (cit. on p. 20).

[72] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses," *Behavior Research Methods*, vol. 41, no. 4, pp. 1149–1160, 2009 (cit. on p. 22).

[73] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 2013 (cit. on p. 22).

[74] P. Gomez, A. von Gunten, and B. Danuser, "Eye gaze behavior during affective picture viewing: Effects of motivational significance, gender, age, and repeated exposure," *Biological Psychology*, vol. 146, p. 107 713, 2019 (cit. on p. 23).

[75] J. Kennedy, P. Baxter, and T. Belpaeme, "Comparing robot embodiments in a guided discovery learning interaction with children," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 293–308, 2015 (cit. on p. 24).

[76] F. Broz, H. Lehmann, C. L. Nehaniv, and K. Dautenhahn, "Automated analysis of mutual gaze in human conversational pairs," in *Eye Gaze in Intelligent User Interfaces*, Springer, 2013, pp. 41–60 (cit. on pp. 27, 41).

[77] H. Lehmann, I. Keller, R. Ahmadzadeh, and F. Broz, "Naturalistic conversational gaze control for humanoid robots-a first step," in *International Conference on Social Robotics*, Springer, 2017, pp. 526–535 (cit. on pp. 27, 56, 57, 61).

[78] C. Acarturk, B. Indurkya, P. Nawrocki, B. Sniezynski, M. Jarosz, and K. A. Usal, "Gaze aversion in conversational settings: An investigation based on mock job interview," *Journal of Eye Movement Research*, vol. 14, no. 1, 2021 (cit. on pp. 27, 41, 56, 57, 61).

[79] A. Papoulis, "Brownian movement and markov processes," *Probability, Random Variables, and Stochastic Processes*, pp. 515–553, 1984 (cit. on p. 27).

[80] T. M. Beasley and R. E. Schumacker, "Multiple regression approach to analyzing contingency tables: Post hoc and planned comparison procedures," *The Journal of Experimental Education*, vol. 64, no. 1, pp. 79–93, 1995 (cit. on p. 28).

[81] M. A. Garcia-Perez and V. Nunez-Anton, "Cellwise residual analysis in two-way contingency tables," *Educational and Psychological Measurement*, vol. 63, no. 5, pp. 825–839, 2003 (cit. on p. 28).

[82] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources," *Psychological Bulletin*, vol. 91, no. 2, pp. 276–292, 1982 (cit. on p. 33).

[83] S. Joshi, Y. Li, R. M. Kalwani, and J. I. Gold, "Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex," *Neuron*, vol. 89, no. 1, pp. 221–234, 2016 (cit. on p. 33).

[84] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 185–198, 2003 (cit. on p. 34).

[85] M. St-Yves, "The psychology of rapport: Five basic rules," *Investigative Interviewing*, pp. 82–106, 2006 (cit. on p. 35).

[86] S. Andrist, W. Collier, M. Gleicher, B. Mutlu, and D. Shaffer, "Look together: Analyzing gaze coordination with epistemic network analysis," *Frontiers in Psychology*, vol. 6, pp. 1–15, 2015 (cit. on p. 40).

[87] A. Clodic, R. Alami, and R. Chatila, "Key elements for joint human-robot action," in *Robo-Philosophy*, IOS Press Ebooks, vol. 273, 2014, pp. 23–33 (cit. on pp. 41, 42, 53, 54, 57, 61).

[88] S. Baron-Cohen, "How to build a baby that can read minds: Cognitive mechanisms in mindreading," *Current Psychology Letters: Behaviour, Brain and Cognition*, 1994 (cit. on pp. 42–44, 49, 103).

[89] S. Baron-Cohen, *Mindblindness: An essay on autism and theory of mind*. MIT Press, 1997 (cit. on pp. 42, 43, 103).

[90] M. Koller, A. Weiss, and M. Vincze, "I see what you did there: Towards a gaze mechanism for joint actions in human-robot interaction," in *Trust in Robots*, M. Vincze and S. Koeszegi, Eds., Vienna: TU Wien Academic Press, 2023, pp. 149–178 (cit. on pp. 42, 103, 104, 106).

[91] N. Akhtar and M. A. Gernsbacher, "Joint attention and vocabulary development: A critical look," *Linguistics and Language Compass*, vol. 1, no. 3, pp. 195–207, 2007 (cit. on p. 43).

[92] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, vol. 1, no. 4, pp. 515–526, 1978 (cit. on p. 43).

[93] S. R. Langton, R. J. Watt, and V. Bruce, "Do the eyes have it? cues to the direction of social attention," *Trends in Cognitive Sciences*, vol. 4, no. 2, pp. 50–59, 2000 (cit. on p. 44).

[94] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, "Understanding and sharing intentions: The origins of cultural cognition," *Behavioral and Brain Sciences*, vol. 28, no. 5, pp. 675–691, 2005 (cit. on p. 45).

[95] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognitive Psychology*, vol. 8, no. 4, pp. 441–480, 1976 (cit. on p. 45).

[96] P. E. Cairns and A. L. Cox, *Research methods for human-computer interaction*. Cambridge University Press, 2008 (cit. on p. 45).

[97] F. Kaplan and V. V. Hafner, "The challenges of joint attention," *Interaction Studies*, vol. 7, no. 2, pp. 135–169, 2006 (cit. on p. 46).

[98] A. Kolbeinsson, E. Lagerstedt, and J. Lindblom, "Foundation for a classification of collaboration levels for human-robot cooperation in manufacturing," *Production and Manufacturing Research*, vol. 7, no. 1, pp. 448–471, 2019 (cit. on p. 46).

[99] M. Imai, T. Ono, and H. Ishiguro, "Physical relation and expression: Joint attention for human-robot interaction," *IEEE Transactions on Industrial Electronics*, vol. 50, no. 4, pp. 636–643, 2003 (cit. on p. 47).

[100] C.-M. Huang and A. L. Thomaz, "Effects of responding to, initiating and ensuring joint attention in human-robot interaction," in *2011 20th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2011, pp. 65–71 (cit. on p. 47).

[101] A. Pereira, C. Oertel, L. Fermoselle, J. Mendelson, and J. Gustafson, "Responsive joint attention in human-robot interaction," *International Conference on Intelligent Robots and Systems*, pp. 1080–1087, 2019 (cit. on p. 48).

[102] R. M. Frankel, M. Flanagan, P. Ebright, *et al.*, "Context, culture and (non-verbal) communication affect handover quality," *BMJ Quality and Safety*, vol. 21, no. Suppl 1, pp. 121–128, 2012 (cit. on p. 48).

[103] E. C. Grigore, K. Eder, A. G. Pipe, C. Melhuish, and U. Leonards, "Joint action understanding improves robot-to-human object handover," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2013, pp. 4622–4629 (cit. on p. 48).

[104] A. Moon, D. M. Troniak, B. Gleeson, *et al.*, "Meet me where i'm gazing: How shared attention gaze affects human-robot handover timing," in *2014 9th ACM/IEEE International Conference on Human-Robot Interaction*, 2014, pp. 334–341 (cit. on p. 48).

[105] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990 (cit. on pp. 49, 65, 67).

[106] S. Coradeschi, A. Loutfi, and B. Wrede, "A short review of symbol grounding in robotic and intelligent systems," *KI-Künstliche Intelligenz*, vol. 27, pp. 129–136, 2013 (cit. on pp. 49, 67).

[107] R. Mirsky, X. Xiao, J. Hart, and P. Stone, "Prevention and resolution of conflicts in social navigation–a survey," *arXiv preprint arXiv:2106.12113*, 2021 (cit. on p. 49).

[108] S. M. LaValle, *Planning algorithms*. Cambridge University Press, 2006 (cit. on pp. 49, 50).

[109] M. Ghallab, D. Nau, and P. Traverso, *Automated planning and acting*. Cambridge University Press, 2016 (cit. on p. 51).

[110] M. Fox and D. Long, "Pddl2.1: An extension to pddl for expressing temporal planning domains," in *Journal of Artificial Intelligence Research*, vol. 20, 2003, pp. 61–124 (cit. on p. 51).

[111] V. Lifschitz, "On the semantics of strips," in *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop*, 1987, pp. 1–9 (cit. on p. 51).

[112] M. Ramírez and H. Geffner, "Probabilistic plan recognition using off-the-shelf classical planners," in *AAAI Conference on Artificial Intelligence*, 2010 (cit. on pp. 51, 64, 66, 69, 73, 74, 87).

[113] G. Sukthankar, C. Geib, H. H. Bui, D. Pynadath, and R. P. Goldman, *Plan, activity, and intent recognition: Theory and practice*. Newnes, 2014 (cit. on pp. 51, 66).

[114] S. Sohrabi, A. V. Riabov, and O. Udrea, "Plan recognition as planning revisited," in *IJCAI International Joint Conference on Artificial Intelligence*, New York, NY, 2016, pp. 3258–3264 (cit. on pp. 51, 66, 74).

[115] M. Johnson, C. M. Jonker, M. B. van Riemsdijk, P. J. Feltovich, and J. M. Bradshaw, "Joint activity testbed: Blocks world for teams (bw4t)," in *International Workshop on Engineering Societies in the Agents World*, vol. 9, 2009, pp. 254–256 (cit. on p. 52).

[116] K. A. Barchard, L. Lapping-Carr, R. S. Westfall, A. Fink-Armold, S. B. Banisetty, and D. Feil-Seifer, "Measuring the perceived social intelligence of robots," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 4, pp. 1–29, 2020 (cit. on p. 52).

[117] A. B. Jensen, "Towards verifying a blocks world for teams goal agent," in *13th International Conference on Agents and Artificial Intelligence*, Science and Technology Publishing, 2021, pp. 337–344 (cit. on p. 52).

[118] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2005, pp. 708–713 (cit. on p. 55).

[119] L. Takayama, D. Dooley, and W. Ju, "Expressing thought: Improving robot readability with animation principles," in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction*, 2011, pp. 69–76 (cit. on p. 55).

[120] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends in Cognitive Sciences*, vol. 9, no. 4, pp. 188–194, 2005 (cit. on p. 55).

[121] J. Pelz, M. Hayhoe, and R. Loeber, "The coordination of eye, head, and hand movements in a natural task," *Experimental Brain Research*, vol. 139, no. 3, pp. 266–277, 2001 (cit. on p. 55).

[122] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," *International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1151–1160, 2014 (cit. on p. 57).

[123] C. Lugaresi, J. Tang, H. Nash, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019 (cit. on p. 59).

[124] M. Beetz, F. Balint-Benczedi, N. Blodow, D. Nyga, T. Wiedemeyer, and Z.-C. Marton, "Robosherlock: Unstructured information processing for robot perception," *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1549–1556, 2015 (cit. on p. 63).

[125] M. Tenorth and M. Beetz, "Representations for robot knowledge in the knowrob framework," *Artificial Intelligence*, vol. 247, pp. 151–169, 2017 (cit. on pp. 63, 67).

[126] A. Zaatri, "Overview of some command modes for human-robot interaction systems," *Journal of Information Systems Engineering and Management*, vol. 7, no. 2, p. 14 039, 2022 (cit. on p. 64).

[127] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022 (cit. on p. 64).

[128] E. E. Aksoy, A. Orhan, and F. Wörgötter, "Semantic Decomposition and Recognition of Long and Complex Manipulation Action Sequences," *International Journal of Computer Vision*, vol. 122, no. 1, pp. 84–115, 2017 (cit. on pp. 65, 67, 68).

[129] E. E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, and F. Wörgötter, "Learning the semantics of object-action relations by observation," *International Journal of Robotics Research*, vol. 30, no. 10, pp. 1229–1249, 2011 (cit. on pp. 65, 67).

[130] S. Yang, Q. Gao, C. Liu, C. Xiong, S.-C. Zhu, and J. Chai, "Grounded semantic role labeling," in *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 149–159 (cit. on pp. 65, 73, 74, 80).

[131] M. Rohrbach, A. Rohrbach, M. Regneri, *et al.*, "Recognizing fine-grained and composite activities using hand-centric features and script data," *International Journal of Computer Vision*, pp. 1–28, 2015 (cit. on pp. 66, 73, 79–81).

[132] M. Koller, T. Patten, and M. Vincze, "Plan recognition from object detection traces," *AAAI Workshop on Plan, Activity, and Intent Recognition (PAIR)*, 2020. [Online]. Available: `http://www.planrec.org/PAIR/PAIR\%2020/Resource_files/PAIR20papers.zip` (cit. on pp. 66, 103, 105, 106).

[133] R. G. Freedman and S. Zilberstein, "A unifying perspective of plan, activity, and intent recognition," *Workshop on Plan, Activity, and Intent Recognition*, pp. 1–8, 2019 (cit. on p. 66).

[134] M. Ramírez and H. Geffner, "Goal recognition over POMDPs: Inferring the intention of a POMDP agent," *IJCAI International Joint Conference on Artificial Intelligence*, pp. 2009–2014, 2011 (cit. on p. 66).

[135] S. Keren, A. Gal, and E. Karpas, "Goal recognition design," *International Conference on Automated Planning and Scheduling*, vol. 24, pp. 154–162, 2014 (cit. on p. 66).

[136] S. Keren, A. Gal, and E. Karpas, "Goal recognition design for non-optimal agents," *National Conference on Artificial Intelligence*, vol. 5, pp. 819–825, 2015 (cit. on p. 66).

[137] R. G. Freedman and S. Zilberstein, "Integration of planning with recognition for responsive interaction using classical planners," *31st AAAI Conference on Artificial Intelligence*, pp. 4581–4588, 2017 (cit. on p. 66).

[138] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, pp. 1–28, 2015 (cit. on p. 67).

[139] S. V. Albrecht and P. Stone, "Autonomous agents modelling other agents: A comprehensive survey and open problems," *Artificial Intelligence*, vol. 258, pp. 66–95, 2018 (cit. on p. 67).

[140] R. Cubek, W. Ertel, and G. Palm, "High-level learning from demonstration with conceptual spaces and subspace clustering," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2015, pp. 2592–2597 (cit. on p. 67).

[141] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017, pp. 4444–4451 (cit. on p. 67).

[142] S. Lemaignan, R. Ros, L. Mösenlechner, R. Alami, and M. Beetz, "ORO, a knowledge management platform for cognitive architectures in robotics," *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems*, pp. 3548–3553, 2010 (cit. on p. 67).

[143] R. L. Granada, R. F. Pereira, J. Monteiro, D. D. A. Ruiz, R. C. Barros, and F. R. Meneguzzi, "Hybrid activity and plan recognition for video streams," in *31st AAAI Conference: Plan, Activity and Intent Recognition Workshop*, 2017 (cit. on pp. 67, 80).

[144] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788 (cit. on p. 68).

[145] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969 (cit. on p. 68).

[146] Y. Jiang, N. Walker, J. Hart, and P. Stone, "Open-world reasoning for service robots," in *International Conference on Automated Planning and Scheduling*, vol. 29, 2019, pp. 725–733 (cit. on p. 69).

[147] A. Saffidine, T. Cazenave, and J. Méhat, "Ucd: Upper confidence bound for rooted directed acyclic graphs," *Knowledge-Based Systems*, vol. 34, pp. 26–33, 2012 (cit. on pp. 71, 72).

[148] R. Howey, D. Long, and M. Fox, "Val: Automatic plan validation, continuous effects and mixed initiative planning using pddl," in *16th IEEE International Conference on Tools with Artificial Intelligence*, IEEE, 2004, pp. 294–301 (cit. on p. 71).

[149] B. Bonet and H. Geffner, "Planning under partial observability by classical replanning: Theory and experiments," in *22nd International Joint Conference on Artificial Intelligence*, 2011, pp. 1936–1941 (cit. on p. 77).

[150] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017 (cit. on p. 79).

[151] A. D. Dragan, A. L. Thomaz, and S. S. Srinivasa, "Collaborative manipulation: New challenges for robotics and hri," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction*, IEEE, 2013, pp. 435–436 (cit. on p. 79).

[152] S. Trick, D. Koert, J. Peters, and C. A. Rothkopf, "Multimodal uncertainty reduction for intention recognition in human-robot interaction," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2019, pp. 7009–7016 (cit. on p. 79).

[153] M Begum, R Huq, R Wang, and A Mihailidis, "Collaboration of an assistive robot and older adults with dementia," *Gerontechnology*, vol. 13, no. 4, pp. 405–419, 2015 (cit. on p. 79).

[154] M. Karg and A. Kirsch, "Acquisition and use of transferable, spatio-temporal plan representations for human-robot interaction," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2012, pp. 5220–5226 (cit. on p. 79).

[155] M. L. Walters, M. A. Oskoei, D. S. Syrdal, and K. Dautenhahn, "A long-term human-robot proxemic study," in *2011 20th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2011, pp. 137–142 (cit. on p. 79).

[156] X. Gao, R. Gong, T. Shu, X. Xie, S. Wang, and S.-C. Zhu, "Vrkitchen: An interactive 3d virtual environment for task-oriented learning," *arXiv preprint arXiv:1903.05757*, 2019 (cit. on pp. 80–82, 84).

[157] B. Shen, F. Xia, C. Li, *et al.*, "Igibson 1.0: A simulation environment for interactive tasks in large realistic scenes," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 7520–7527 (cit. on p. 80).

[158] K. Dimitropoulos, I. Hatzilygeroudis, and K. Chatzilygeroudis, "A brief survey of sim2real methods for robot learning," in *International Conference on Robotics in Alpe-Adria Danube Region*, Springer, 2022, pp. 133–140 (cit. on p. 80).

[159] A. A. Rusu, M. Vecerik, T. Rothörl, N. Heess, R. Pascanu, and R. Hadsell, "Sim-to-real robot learning from pixels with progressive nets," *arXiv preprint arXiv:1610.04286*, 2016 (cit. on p. 80).

[160] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009 (cit. on p. 80).

[161] D. Damen, H. Doughty, G. Maria Farinella, *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision*, 2018, pp. 720–736 (cit. on p. 81).

[162] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2847–2854, 2012 (cit. on p. 81).

[163] A. Garcia-Garcia, P. Martinez-Gonzalez, S. Oprea, *et al.*, "The robotrix: An extremely photorealistic and very-large-scale indoor dataset of sequences with robot trajectories and interactions," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 6790–6797 (cit. on p. 81).

[164] X. Puig, K. Ra, M. Boben, *et al.*, "Virtualhome: Simulating household activities via programs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8494–8502 (cit. on p. 81).

[165] C. Li, F. Xia, R. Martín-Martín, *et al.*, "Igibson 2.0: Object-centric simulation for robot learning of everyday household tasks," *arXiv preprint arXiv:2108.03272*, 2021 (cit. on p. 81).

[166] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "Rgb-d-based action recognition datasets: A survey," *Pattern Recognition*, vol. 60, pp. 86–105, 2016 (cit. on p. 81).

[167] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013 (cit. on p. 81).

[168] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013 (cit. on p. 81).

[169] M. Koller, T. Patten, and M. Vincze, "A new vr kitchen environment for recording well annotated object interaction tasks," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction*, IEEE, 2022, pp. 629–633 (cit. on pp. 81, 103, 106).

[170] *Virtual annotated kitchen environment dataset*, TU Data, doi: *10.48436/r5d7q-bdn48*, 2021 (cit. on p. 82).

[171] *Virtual annotated kitchen environment dataset - Ego erspective*, TU Data, doi: *10.48436/9y2x1-q4n71*, 2021 (cit. on p. 82).

[172] A. Shimada, K. Kondo, D. Deguchi, G. Morin, and H. Stern, "Kitchen scene context based gesture recognition: A contest in icpr2012," in *International Workshop on Depth Image Analysis and Applications*, Springer, 2012, pp. 168–185 (cit. on p. 86).

[173] G. Ainslie and N. Haslam, "Hyperbolic discounting," in *G. Loewenstein & J. Elster (Eds.), Choice Over Time*, Russell Sage Foundation, 1992, pp. 57–92 (cit. on p. 87).

[174] N. Schwarz, H. Bless, F. Strack, G. Klumpp, H. Rittenauer-Schatka, and A. Simons, "Ease of retrieval as information: Another look at the availability heuristic," *Journal of Personality and Social psychology*, vol. 61, no. 2, pp. 195–202, 1991 (cit. on p. 87).

[175] L. Festinger, "Cognitive dissonance," *Scientific American*, vol. 207, no. 4, pp. 93–106, 1962 (cit. on p. 87).

[176] D. Kahneman, "Prospect theory: An analysis of decisions under risk," *Econometrica*, vol. 47, pp. 363–391, 1979 (cit. on pp. 87, 90).

[177] M. Kwon, E. Biyik, A. Talati, K. Bhasin, D. P. Losey, and D. Sadigh, "When humans aren't optimal: Robots that collaborate with risk-aware humans," in *2020 15th ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 43–52 (cit. on pp. 88–90, 92, 93).

[178] S. Milli, D. Hadfield-Menell, A. Dragan, and S. Russell, "Should robots be obedient?" *arXiv preprint arXiv:1705.09990*, 2017 (cit. on pp. 88, 89, 91).

[179] O. Evans, A. Stuhlmüller, and N. Goodman, "Learning the preferences of ignorant, inconsistent agents," in *AAAI Conference on Artificial Intelligence*, vol. 30, 2016, pp. 323–329 (cit. on pp. 89, 91).

[180] D. P. Losey and D. Sadigh, "Robots that take advantage of human trust," *arXiv preprint arXiv:1909.05777*, 2019 (cit. on pp. 89, 91).

[181] H. Fennema and P. Wakker, "Original and cumulative prospect theory: A discussion of empirical differences," *Journal of Behavioral Decision Making*, vol. 10, no. 1, pp. 53–64, 1997 (cit. on pp. 90, 98).

[182] M. Koller, T. Patten, and M. Vincze, "Risk-averse biased human policies in assistive multi-armed bandit settings," *TRAITS Workshop, 16th ACM/IEEE International Conference on Human-Robot Interaction, arXiv preprint arXiv:2104.05334*, 2021 (cit. on pp. 90, 103, 105, 106).

[183] M. Koller, T. Patten, and M. Vincze, "Risk-averse biased human policies with a robot assistant in multi-armed bandit settings," in *The 14th PErvasive Technologies Related to Assistive Environments Conference*, 2021, pp. 483–488 (cit. on pp. 90, 103, 105, 106).

[184] S. Vakili and Q. Zhao, "Mean-variance and value at risk in multi-armed bandit problems," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2015, pp. 1330–1335 (cit. on p. 90).

[185] N. Galichet, M. Sebag, and O. Teytaud, "Exploration vs exploitation vs safety: Risk-aware multi-armed bandits," in *Asian Conference on Machine Learning*, PMLR, 2013, pp. 245–260 (cit. on p. 90).

[186] O.-A. Maillard, "Robust risk-averse stochastic multi-armed bandits," in *International Conference on Algorithmic Learning Theory*, Springer, 2013, pp. 218–233 (cit. on p. 90).

[187] A. Cassel, S. Mannor, and A. Zeevi, "A general approach to multi-armed bandits under risk criteria," in *Conference On Learning Theory*, PMLR, 2018, pp. 1295–1306 (cit. on p. 90).

[188] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "The off-switch game," *arXiv preprint arXiv:1611.08219*, 2016 (cit. on p. 91).

[189] M. O. Rieger, M. Wang, and T. Hens, "Estimating cumulative prospect theory parameters from an international survey," *Theory and Decision*, vol. 82, no. 4, pp. 567–596, 2017 (cit. on p. 96).

[190] A. M. Miron and J. W. Brehm, "Reactance theory-40 years later," *Zeitschrift für Sozialpsychologie*, vol. 37, no. 1, pp. 9–18, 2006 (cit. on p. 99).

[191] S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," *Ai Magazine*, vol. 36, no. 4, pp. 105–114, 2015 (cit. on pp. 103, 105).

[192] F. Heider and M. Simmel, "An experimental study of apparent behavior," *The American Journal of Psychology*, vol. 57, no. 2, pp. 243–259, 1944 (cit. on p. 104).

[193] D. C. Dennett, *The intentional stance*. MIT Press, 1987 (cit. on pp. 104, 105).

Erklärung

Hiermit erkläre ich, dass die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt wurde. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder in ähnlicher Form in anderen Prüfungsverfahren vorgelegt.

Vienna, February 2023

Mag.rer.nat. Michael Koller B.Sc.